An Out-of-the-Lab Evaluation of Dry EEG Technology on a Large-Scale Motor Imagery Brain-Computer Interface Dataset

An Out-of-the-Lab Evaluation of Dry EEG Technology on a Large-Scale Motor Imagery Brain-Computer Interface Dataset

M Sultana[1], A Matran-Fernandez[1], S Halder[1], R Nawaz[1], O Jain[2], R Scherer[1], R Chavarriaga[3], JdR Millán[4,5,6] and S Perdikis[1]

[1] Brain-Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK
[2] Computing and Psychology Department, Goldsmiths, University of London, London, UK
[3] Responsible AI Innovation Group, ZHAW School of Engineering, Winterthur, Switzerland
[4] Chandra Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78712, US
[5] Department of Neurology, The University of Texas at Austin, Austin, Texas 78712, US
[6] Department of Biomedical Engineering, The University of Texas at Austin, Austin, Texas 78712, US

E-mail: serafeim.perdikis@essex.ac.uk

August 2025

Abstract.

Objective. This study assesses the signal quality of state-of-the-art dry electroencephalography (EEG) under highly challenging, uncontrolled, real-world conditions and compares it to conventional wet EEG. *Approach.* EEG data from 530 participants recorded during a public exhibition were benchmarked against several established signal quality metrics, including spiking activity, kurtosis, Auto-Mutual Information (AMI), spectral entropy, gamma-band power, and parameters extracted using the Fitting Oscillations and One-Over F (FOOF) model. Additionally, ICLabel decomposition was applied to quantify artifact influences across EEG channels. Dry electrode results were compared with their equivalents extracted on two control datasets comprising 71 and 80 participants, respectively, recorded with wet EEG systems in laboratory, home, or clinical surroundings. *Main Results* The analysis revealed condition-specific susceptibility to artifacts for both EEG modalities. The dry EEG system exhibited substantial robustness in moderate-noise scenarios, with artifact profiles comparable to controlled wet EEG recordings. However, recordings obtained in highly dynamic conditions showed increased muscle artifacts and broadband activity, notably in frontal and temporal regions. Wet EEG systems, under controlled conditions, were overall less inflicted by artifacts, yet, fronto-central ocular and muscular artifacts were consistently present. ICLabel analysis further confirmed these findings, indicating similar proportions of brain-related activity across systems (approximately 31–49.5%), but highlighted increased vulnerability to movement and environmental artifacts in dry EEG during dynamic tasks. *Significance.* In agreement with recent similar investigations, our findings demonstrate that dry EEG caps have significantly matured, achieving signal quality comparable to wet EEG systems even in challenging real-world conditions, provided appropriate artifact mitigation strategies are employed. These results affirm the practical readiness and broad feasibility of dry EEG technologies for diverse Brain-Computer Interface (BCI) applications in naturalistic environments.

Keywords: electroencephalography, EEG, dry EEG, wet EEG, signal quality, artifacts, benchmarking, motor imagery, brain-computer interface

1. Introduction

Non-invasive EEG continues to be the most widely used neuroimaging modality for BCI applications, owing to its portability, cost-effectiveness, and relatively non-obtrusive nature. Traditionally, high-quality EEG recordings are achieved through gel-based ("wet") electrodes, which significantly enhance signal conductivity by applying gel in between the scalp and the electrodes' surface to reduce the impedance of the skin-electrode interface. Despite their superior Signal-to-Noise Ratio (SNR), wet EEG systems face substantial practical limitations on account of the gel application necessity, including prolonged setup times, availability of an expert to apply the gel, potential user discomfort, need for personal hygiene after use, and difficulty in maintaining stable signal quality during extended recording sessions [1].

In response to these challenges, dry EEG technology has gained significant attention over the past 15 years. These systems, characterized by electrodes that do not require conductive gel or saline solutions, allow for faster and autonomous preparation, enhanced portability, significantly improved user comfort, and minimal, seamless overall donning and doffing procedures, thereby facilitating the application of EEG-based BCIs beyond strictly controlled laboratory conditions [2]. Such advantages are critical for broadening EEG-based BCI applications into everyday, real-world environments, including consumer health monitoring, mobile health applications (assistive technology, rehabilitation, etc.), as well as several non-health-related, consumer-oriented applications (gaming, neuromarketing, and others).

However, initial evaluations of dry EEG technologies reported several notable challenges, such as reduced signal quality, higher impedance, and susceptibility to motion artifacts [2]. Despite these early issues, advancements in sensor materials, electrode designs, and adaptive signal processing techniques have significantly improved their performance, making dry EEG systems increasingly viable for practical BCI implementations. Some comparative studies [3, 1], have shown that current dry EEG technologies are capable of reliably capturing event-related potentials (ERPs) and other essential EEG features. Guger et al.[4] reported comparable accuracy between dry and wet electrodes in P300-based BCI spelling tasks, supporting the practical usability of dry EEG systems in realistic scenarios. Different studies [2, 5, 6] further underscored these findings by demonstrating dry electrode reliability in capturing clinically relevant EEG signals.

Recent research has increasingly focused on evaluating dry EEG systems specifically within Sensorimotor Rhythms (SMR) paradigms [7, 8, 9]. For example, one study conducted simultaneous recordings using dry and wet electrodes during motor imagery tasks, revealing that despite higher impedance and susceptibility to noise, dry electrodes could effectively capture SMR features critical for successful BCI operation [7]. Another investigation comparing active dry electrodes to active and passive wet electrodes in EEG signal quality reported that active dry systems provided comparable signal quality and temporal stability, suggesting their suitability for practical EEG applications, including those demanding stable and prolonged monitoring [9]. A further comparative analysis of clinical EEG recordings found that dry electrodes, despite a modest SNR reduction, yielded statistically equivalent signal quality indices—encompassing overall waveform characteristics, artifact metrics and spectral noise levels—compared to wet electrodes; with standard preprocessing, the dry system supported sound clinical interpretations [10]. A comprehensive review additionally highlighted advancements in dry electrode technology, emphasizing improvements in electrode impedance, signal reliability, and practical usability, affirming that state-of-the-art dry EEG systems are increasingly competitive with traditional wet EEG setups across multiple BCI paradigms [11].

The need for dry EEG systems is especially compelling for long-term, continuous brain activity monitoring scenarios, such as seizure prediction in epilepsy patients [12, 13], attention training interventions for ADHD [14], driver fatigue monitoring [15], and cognitive workload assessments [16]. These potential applications emphasize the added value that dry EEG systems can bring to the neurotechnology market, provided that their quality is proven to be adequate.

While this literature has provided valuable information on the competitiveness of dry EEG technology, these investigations have been limited by small participant numbers and confined to controlled laboratory conditions, leaving uncertainties regarding the practical usability of dry EEG out of the lab. Consequently, there remains a substantial need for well-powered, comprehensive validation of dry EEG technologies, especially considering environments characterized by high noise levels and likelihood of artifact contamination.

More elaborately, electrical potentials recorded from the scalp approximately lie within the rather minuscule [-150, +150] µV range, making EEG signals particularly vulnerable to contamination from physiological artifacts—such as ocular movements, muscle activity, cardiac signals, and head movements [17, 18, 19, 20, 21]—as well as non-physiological interference, including electromagnetic disturbances, technical malfunctions, and external physical disruptions [19, 22]. Although robust EEG recording systems and standard-

ized experimental protocols can significantly mitigate these artifacts, completely avoiding contamination remains practically impossible, especially in real-world and uncontrolled environments. EEG data collected outside laboratory conditions—with both wet and dry EEG systems—often exhibit decreased SNR due to persistent artifact contamination. While most modern EEG devices, including those employed in this study, include basic onboard artifact mitigation features, the presence of residual noise is unavoidable. Here, we seek to discern the extent to which the two compared EEG sensor technologies (dry and wet) suffer contamination. Furthermore, we determine and juxtapose the reliability of the Motor Imagery (MI) correlates extracted from the respective recordings.

To address these gaps in dry sensor assessment, we evaluate the signal quality of a commercial dry EEG solution in an unprecedentedly large cohort of 530 participants. Our analysis uses a broad set of direct and indirect metrics for EEG quality and the dry system is compared to two widely used, research-grade, wet-EEG systems. Furthermore, the dry recordings were collected at a public art/science exhibition called "Mental Work" [23], held in a large, very noisy public exhibition hall with no particular space adaptations to accommodate the use of EEG in its premises. This study has been largely motivated by early (at this stage, qualitative and anecdotal) signs during Mental Work that, in spite of dry electrode signals having been—and still largely are—considered a priori inferior to wet ones, and although dry EEG data in this case were collected in highly noisy, uncontrolled surroundings, the quality of the extracted EEG signals acquired seemed largely comparable to that of traditional wet-electrode systems. This article attempts to put these impressions to the test, providing a quantified, principled, reliable and precise comparison of dry and wet EEG exploiting the availability of big MI BCI datasets for both EEG modalities.

What makes this work truly novel is reporting on dry EEG readiness simultaneously with a very large sample size and within a very challenging real-world context. Unlike all prior dry-vs-wet EEG comparisons—which, as already mentioned, have been either conducted in tightly controlled laboratory settings or on small cohorts, or both—this work is the first to evaluate EEG signals in a realistic setting substantially "polluted" with noise and with a number of participants that is more than an order of magnitude greater than the field's norms, thus greatly improving the reliability of findings and informing on the maturity of this technology to be deployed to realistic scenarios. We focus on generic, paradigm-independent EEG signal quality metrics, but additionally report open-loop MI classification accuracy exploiting the underlying paradigm of the Mental Work exhibition. Of note, SMR BCI, being endogenous (thus, particularly susceptible to user-generated, physiological artifacts) and inherently low SNR, offers a very stringent stress-test for electrode robustness, perhaps more so than stimuli-driven, Event-Related Potential (ERP)-based paradigms such as P300 and Steady-state Visually Evoked Potentials (SSVEPs) on which the majority of the relevant literature has focused. We further aim to investigate whether standard artifact-detection and removal pipelines remain valid on these extremely noisy, real-world recordings and try to identify task- and environment-specific artifact fingerprints—such as fronto-temporal muscle bursts that hardly ever emerge in the lab. These insights go well beyond a simplistic conclusion such as "dry equals wet", showing instead how and where dry caps may fail or succeed in real-life applications.

In summary, we hereby aim to determine whether state-of-the-art dry EEG systems have matured sufficiently to deliver signal quality comparable to wet EEG systems under challenging real-world conditions. Our comparative analysis provides critical insights into the readiness of dry EEG caps for broad practical deployment and their potential as reliable alternatives to traditional wet EEG technology in diverse BCI applications.

2. Materials and Methods

2.1. Datasets

We analyze five datasets derived from three large SMR BCI databases, referenced consistently as *Dry Training*, *Dry Control*, *Wet Healthy 1*, *Wet AT 1*, and *Wet Healthy 2*. The main characteristics of all datasets (participants, montage/reference, sampling rate, hardware, environment) are summarized in Table 1 and Fig. 1.

The Dry Training and Dry Control datasets were obtained from a database comprising a total of 530 participants who took part in a public event titled "Mental Work". This exhibition occurred over several months between 2017 and 2018 at the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland [23]. Visitors registered online to attend scheduled BCI sessions, during which they wore dry-electrode EEG headsets to operate a set of machines inspired by the industrial revolution by means of a two-class MI-based BCI control system (Fig. 1(a)).

Another two datasets, namely, Wet Healthy 1 and Wet AT 1, were acquired with g.USBamp wet systems from $N = 46$ able-bodied volunteers (Wet Healthy 1) and $N = 25$ individuals with

Table 1. Summary of datasets: participants, demographics, channels, sampling, hardware, and recording environment.

Dataset	Participants	Demographics (mean age ± SD)	Channels / Reference	Sampling	Hardware	Environment / Notes
Dry Training	$N = 530$ visitors (public exhibition)[23]	-	**19 dry (10–20)**: Fp1, Fp2, Fz, F3, F4, F7, F8, Cz, C3, C4, C5, C6, T7, T8, Pz, P3, P4, O1, O2; ref = avg ear-lobes(Fig. 1(b1))	300 Hz; no hardware filter	DSI-24 (Wearable Sensing, San Diego, USA)	Semi-controlled booth inside exhibition hall; open-loop MI calibration
Dry Control	same N as above	-	Same as Dry Training	Same as Dry Training	Same as Dry Training	In-the-wild machine control in large public space (crowd, movement, EMI); close-loop machine control
Wet Healthy 1	$N = 46$ able-bodied (EPFL)[24]	41 ± 9 y; F/M: 10/36	**16 wet (10–20)**: Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4; ref = right ear-lobe(Fig. 1(b2))	512 Hz with a hardware band-pass filter (cut-offs at 0.1 Hz and 100 Hz) and notch-filter (50 Hz)	g.USBamp (g.Tec, Austria)	Controlled laboratory; calibration and online MI runs
Wet AT 1	$N = 25$ AT users (SUVA/home)[24, 25]	56 ± 26 y; F/M: 3/22	Same as Wet Healthy 1	Same as Wet Healthy 1	Same as Wet Healthy 1	Clinical/home settings; less controlled; MI training for AT
Wet Healthy 2	$N = 80$ healthy novices[26]	29.9 ± 11.5 y; F/M: 41/39	**119 wet (extended 10–20)**; ref = na-sion(Fig. 1(b3)); *subset of channels used in analysis*	1000 Hz with a hardware band-pass filter (cut-offs at 0.05 Hz and 200 Hz)	BrainAmp DC (Brain Products, Germany)	Controlled laboratory; publicly available dataset

motor disabilities (Wet AT 1) undergoing MI-BCI training as part of BCI research activities in EPFL; both used the same 16-channel sensorimotor montage with right-earlobe reference (Fig. 1b2). Participants of the Wet AT 1 dataset presented with various neurological conditions, including myopathy, spinal cord injury, amputation, spinocerebellar ataxia, and multiple sclerosis [25]. The Wet Healthy 1 dataset was recorded in a controlled laboratory environment at EPFL, Switzerland. In contrast, the Wet AT 1 dataset was collected in clinical or home settings, mainly at the SUVA rehabilitation clinic in Sion, Switzerland, and at the participants' own residences, under somewhat uncontrolled conditions, exposed to many more noise sources compared to the laboratory recordings. Nevertheless, these environments were still notably less dynamic and noisy compared to those encountered by the subjects during the Mental Work exhibition. Core acquisition parameters are listed in Table 1.

Lastly, an additional, publicly available high-density SMR BCI dataset is used and referred to as Wet Healthy 2. This includes $N = 80$ healthy novices with no reported neurological disorders recorded with 119 electrodes (BrainAmp DC Amplifier) in the lab [26]. For cross-dataset comparability, analyses use only channels common to the Dry and Wet System 1 layouts (Table 1)

2.2. Ethics and Approval

This study includes three different databases. The Mental Work data collection was approved by the Cantonal Committee of Vaud (VD, Switzerland) for Ethics in Human Research (CER-VD) under protocol number 2017-01746 [23]. The able-bodied user and AT patient data collection was also approved by CER-VD under protocol number PB_2017–00295 (020-15 CCVEM) and earlier versions of it [25]. The collection of Wet Healthy 2 data was approved by the local

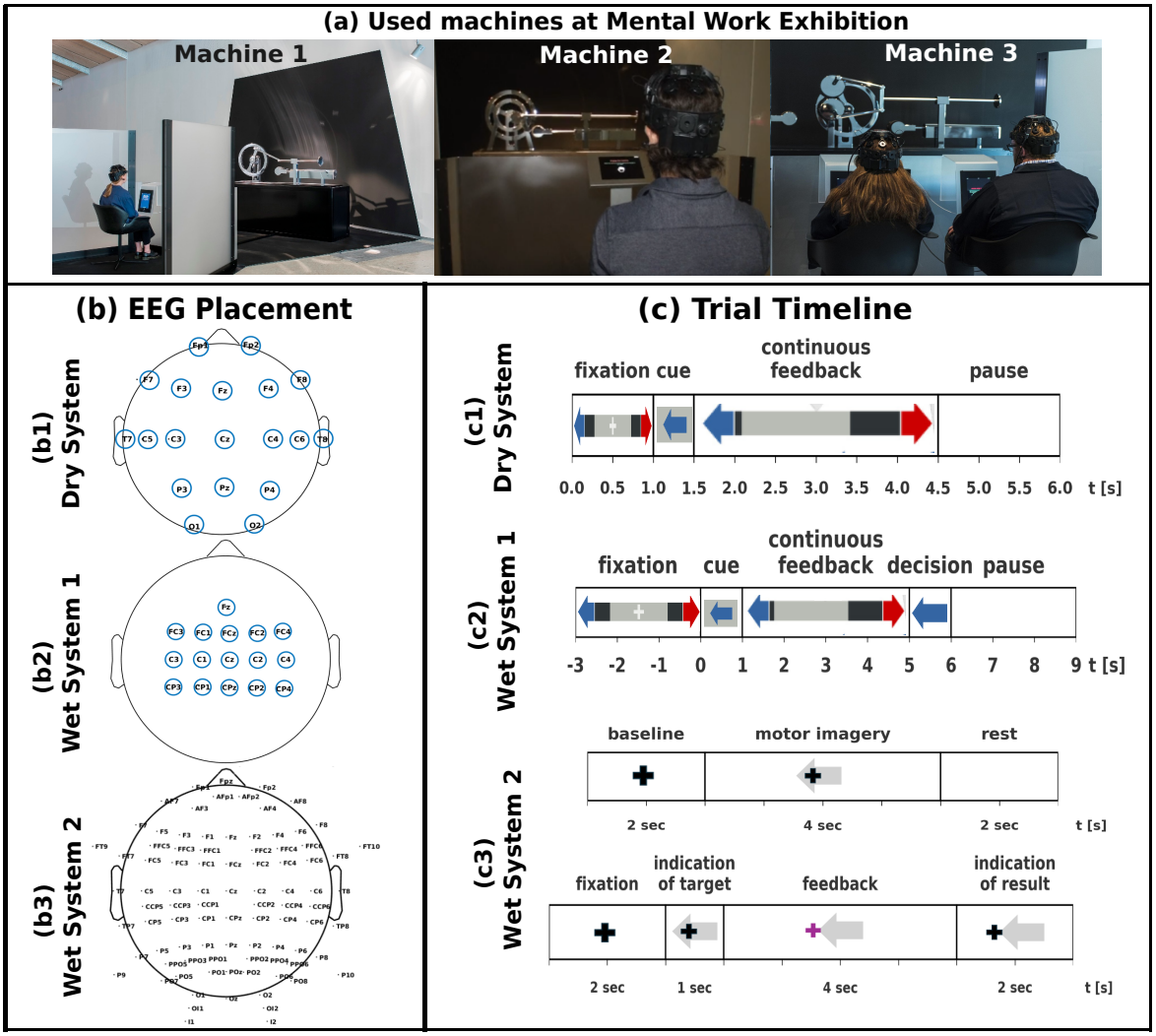

Figure 1. Experimental apparatus and protocol of the three EEG systems. (a) Brain-actuated machines of Mental Work exhibition (b) EEG channel layout and (c) trial timeline of the Dry system (b1, c1), Wet system 1 (b2, c2) and Wet system 2 (b3, c3).

ethics authority (Ethical Review Board of the Medical Faculty of the University of Tübingen) [27]. Informed consent was received from all human subjects, and all experimental protocols were fully compliant with the Declaration of Helsinki and in accordance with local statutory requirements.

2.3. Experimental protocol

After registering with Mental Work, each participant spent roughly 30 minutes collecting data for the calibration of a binary (two-class) MI decoder. Calibration took place inside a small, dedicated, semi-controlled space within the exhibition hall, separated from it by temporary thin walls, thus offering some additional privacy and quietness compared to the large main space devoted to machine control. During calibration, subjects performed 30 cued, 4 s-long trials per MI class (i.e., kinaesthetic imagination of left- and right-hand, or foot movement), in blocks of three runs, yielding a total of 60 MI training trials for each participant (Fig. 1(c1)). After the MI BCI calibration, always in the dedicated booth, subjects proceeded with 2-3 minutes of closed-loop control of a basic feedback graphical user interface displaying a moving visualization bar in real-time, which users attempted to move left or right using MI. These online runs often only contain 4-5 trials and were, as a result, excluded from analysis. Subsequently, participants moved to the large exhibition space to engage with each of the three machines controlled in closed-loop with the participant-specific MI BCI model automatically trained during the calibration phase. It must be highlighted that, unlike a controlled lab environment, the training booth was located within, and loosely separated from, the busy, open-access exhibition hall with abundant foot traffic, background conversation,

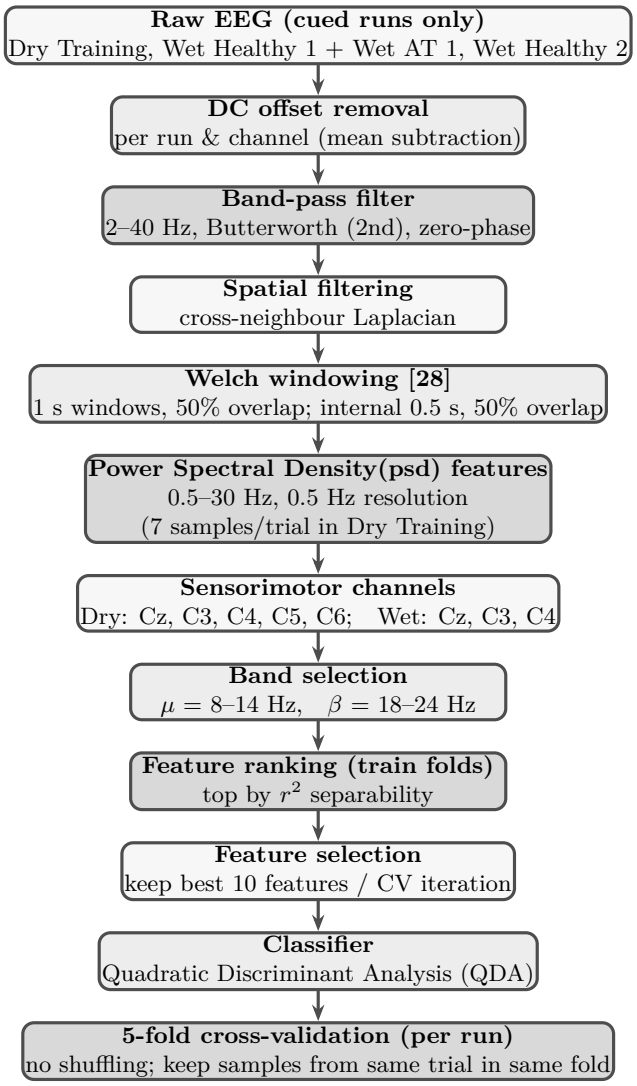

Figure 2. Classification accuracy pipeline (MI decoding).

and electromagnetic interference from phones, lighting, and other electronics.

During machine control, the participants of the Mental Work Exhibition (referred to as "operators") wore the same dry EEG helmet, continuously recording electrical brain activity in a wireless fashion. At the beginning of the phase, participants sat facing a two-meter-long mechanical setup (Machine 1, Fig. 1(a)), consisting of a piston, a flywheel, and a horizontal shaft. Participants use MI-based commands to move the piston towards the flywheel, initiating rotation. This rotation subsequently drove the shaft through a bolt. This interaction was mediated by the trained MI BCI algorithm following the similar setups of previous work [25, 29] but adapted to the specific channel layout of the dry EEG headset. Detection of one class (i.e., when the posterior probability of this mental class given the extracted and selected SMR features would exceed a user-adjustable confidence threshold—usually varying between 70%–90%) would set the piston in motion, and the other class would have no effect. After controlling the machine for some time, BCI control introduced additional complexity, involving more sophisticated machines (Machine 2 and Machine 3, Fig. 1(a)) and task structures. Machines 2 and 3 require two participants to be controlled. They were assigned roles as either "drivers" or "supervisors". Supervisors, through their own MI, could dynamically adjust the probability thresholds, thereby making the drivers' task of controlling the machinery easier or more challenging, and less predictable. Alternatively, supervisors could instruct the BCI of the operator, again via MI, to stop using imagery-based decoding and instead switch to an alpha-wave-based algorithm. In this scenario, supervisors also used MI to cue drivers to relax and clear their minds entirely.

The datasets for both able-bodied participants (Wet Healthy 1) and AT (Wet AT 1) users consist of 1 to 10 MI BCI sessions per subject. Each session includes 3 to 4 calibration (open-loop) and/or online (closed-loop) runs, with each run comprising 15 trials for each MI task. The MI tasks recorded in the database include right-hand, left-hand, both-hand, both-feet MI, and a resting (idling) condition. The experimental protocol is described in detail in [24, 25]. The duration of continuous feedback is fixed at 4 seconds for offline runs (Fig. 1(c2), whereas for online runs, it varies between 2 and 8 seconds depending on a subject-specific timeout.

The experimental protocol consisted of several calibration and feedback runs for the Wet Healthy 2 dataset. During calibration runs, participants performed imagined movements of the left hand, right hand, or feet guided by visual cues (arrows). Each MI task was presented in randomized order across trials, with each trial comprising a 2-second fixation cross epoch, a 4-second imagery period cued throughout by an arrow, and, ultimately, a 2-second pause (inter-trial interval). Each run consisted of 25 trials per MI task, resulting in a total of 225 trials per participant. Feedback runs involved online BCI control based on real-time EEG processing, where participants received visual feedback of their motor imagery performance through cursor movements on the screen. Each feedback trial lasted 9 seconds, comprising an initial 2-second fixation epoch, a 1-second directional cue presentation epoch, then 4 seconds of continuous cursor feedback followed by a 2-second indication of final BCI decision (Fig. 1(c3)).

A detailed description of the experimental setup and protocol generating each of the three datasets can be found in [24, 25, 26]. It is critical to observe that, aside from minor differences in timing and graphics, the

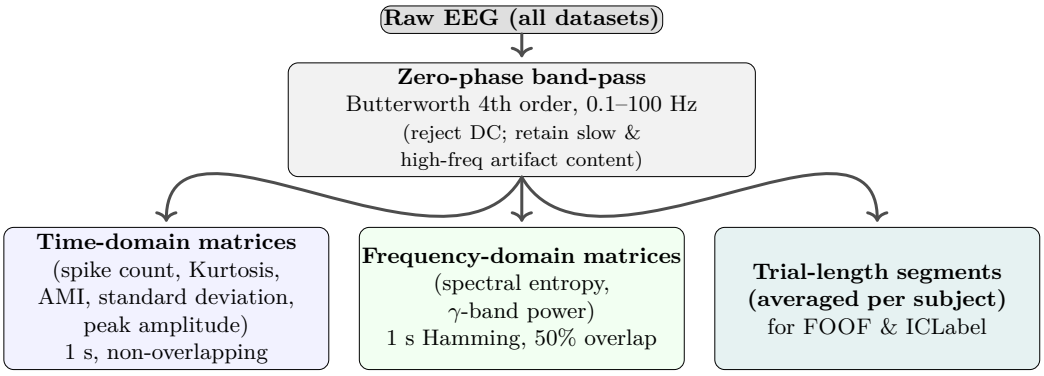

Figure 3. Preprocessing pipeline for direct EEG quality assessment.

experimental protocol of all three datasets is identical. For the purpose of direct comparisons across different datasets, we retained only those channels in Wet Healthy 2 that were common to both Dry EEG (Dry Training, Dry Control) and Wet EEG System 1 (Wet Healthy 1, Wet AT 1).

2.4. Indirect EEG quality assessment

The bulk of our benchmarking analysis focuses on paradigm-independent metrics previously introduced in the literature to directly assess the quality of EEG signals, as elaborated below. However, exploiting the fact that the available datasets have been derived in a BCI context and contain labeled EEG during the execution of motor tasks, we seize the opportunity to also offer indirect evidence of comparative EEG quality through open-loop MI BCI performance and, specifically, by means of the popular and easily interpretable classification accuracy measure.

Classification accuracy: A conventional MI BCI processing pipeline is applied to all datasets to ensure fair cross-dataset comparisons (Fig. 2). Only cued runs that allow extraction of classification accuracy are employed, effectively excluding the Dry Control data of the dry EEG Mental Work dataset. Classification accuracy was computed individually for each run and then averaged per subject. We report both the average accuracy across all runs and the maximum accuracy (i.e., best run) achieved by each participant.

Each participant's classification accuracy was compared against a subject-specific chance level $p_{\text{chance}}(n_i) = 0.5 + \frac{0.506}{\sqrt{n_i}}$ (two-class, binomial distribution of classification decisions assumed following Müller et al. [30]), where n_i is that participant's total number of trials (pooled across runs/sessions). This evaluates to $\approx 58\%$ at $n_i = 40$ and adapts with n_i (higher for fewer trials, lower for more); a participant was considered above chance if $\text{Acc}_i \geq p_{\text{chance}}(n_i)$.

2.5. Direct EEG quality assessment

We additionally sought to systematically quantify the degree to which various common artifacts—such as ocular, muscular, and environmental noise—affected the signal integrity, thus directly evaluating the signal quality that can be delivered by dry and wet EEG technologies, as well as the comparative vulnerability of these to noise. Towards this goal, several time-domain and frequency-domain metrics were used. Prior to the computation of the quality indices, raw EEG data were pre-processed as shown in Fig. 3. No artifact removal or impedance-compensation was applied. Independent Component Analysis (ICA) and ICLabel were used strictly to quantify artifact composition (not to clean the signal), so that residual noise remains measurable across datasets. It is important to emphasize that subject-wise single-sample values are inputted in ANOVA and other statistical testing (and are further averaged per EEG system or per dataset for comparative reporting), so that the data independence assumption is satisfied.

Each of these metrics assesses specific characteristics of the signal, allowing us to evaluate different aspects of EEG signal quality. The battery of signal quality measures applied here has been based on previous work [31, 32, 33, 34] where they have been successfully employed as part of advanced, state-of-the-art artifact removal methods (mainly to distinguish artifact-contaminated from clean independent EEG components). As such, they are widely used and validated metrics for assessing EEG signal quality. The detailed methods employed for each quality metric are outlined below.

2.5.1. Time-Domain Quality Metrics

Spiking activity: Spiking activity quantifies brief but large-magnitude transients mainly associated with muscular artifacts or sudden electrode movements [33]. The presence of spiking activity within EEG segments was quantified using a threshold-based method.

Specifically, for a given EEG channel amplitude $x(t)$, a threshold T was computed dynamically as:
$T = \mu + k\sigma$, where μ and σ represent the mean and standard deviation of the signal segment, respectively. A threshold level of five standard deviations (k=5) above the mean was selected to robustly discriminate high-amplitude transient artifacts from physiological EEG signals, consistent with previously validated artifact detection methodologies in EEG literature [35, 33, 36, 37]. This stringent threshold ensures minimal false-positive detection, focusing predominantly on muscular or movement-related EEG spikes. Samples exceeding this threshold indicate high-magnitude spikes or transient artifacts.

Kurtosis: Kurtosis κ (Eq. 1) was computed to assess the amplitude distribution characteristics of EEG signals over time [38]. Kurtosis κ quantifies the degree to which data distributions are peaked or heavy-tailed, helping to identify transient, high-amplitude artifacts that differ significantly from regular EEG activity. It is defined as

$$\kappa = \frac{\frac{1}{N}\sum_{t=1}^{N}[x(t)-\mu]^4}{\sigma^4} \tag{1}$$

where $x(t)$ is the EEG amplitude at time t, N denotes the number of samples within each 1-second window, μ the mean amplitude, and σ the standard deviation. Kurtosis values significantly higher than that of a Gaussian distribution ($\kappa > 3$) usually indicate abnormal data segments potentially contaminated by artifacts such as muscle bursts or electrode pops [31].

AMI: AMI (Eq. 2) measures nonlinear temporal dependencies within the EEG data, serving as a robust alternative to conventional autocorrelation measures. AMI values outside normal ranges ($\tau = 100$ ms for clean EEG falls in the range of 0.15-0.40 bits) may reflect abnormal temporal dynamics indicative of artifact contamination [32]. For EEG signals represented as discrete random variables X (original) and Y (time-shifted by lag τ), AMI is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

where $p(x,y)$ denotes the joint probability distribution, and $p(x), p(y)$ represent marginal distributions. Here, we used a lag offset of $\tau = 100$ ms, as mentioned in [31], where this choice is based on the observed temporal dynamics of artifacts.

Standard Deviation: The standard deviation σ (Eq. 3, unbiased estimate) was computed to detect channels exhibiting unusually high amplitude variability. This measure captures the magnitude of signal fluctuations around the mean amplitude and can indicate various artifacts, such as muscle contractions (EMG artifacts), electrode displacement, motion-induced noise or loose electrode contacts [31, 22]. If $x(t)$ represents the amplitude at sample t, and μ denotes the mean amplitude of the segment, σ denotes the standard deviation for an EEG segment of length N samples.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}[x(t)-\mu]^2} \tag{3}$$

We identified abnormal channels by applying a robust threshold employing Median Absolute Deviation (MAD), defined as $\sigma_{th} = median(\sigma') + s_{robust}\cdot mad(\sigma')$ and $s_{robust} = 3$ (3 standard deviations from the median), where σ' here is the distribution of standard deviation values retrieved by all 1-sec windows. This method is less affected by outliers and ensures sensitivity to subtle, but consistent deviations indicative of artifacts [39].

Peak Amplitude: Peak amplitude for each EEG segment was defined as $PeakAmplitude = \max_t |x(t)|$. Very high peak amplitudes indicate abrupt transient artifacts such as strong muscular activity (e.g., jaw movements) or mechanical forces applied to electrodes (e.g., pops, presses, etc.) [40].

2.5.2. Frequency-Domain Quality Metrics

Spectral Entropy: Spectral entropy H (Eq. 4) quantifies the uniformity of the Power Spectral Density (PSD), calculated from the normalized PSD, $p(f)$. Higher entropy values indicate a broader and more evenly distributed spectral content, often associated with broadband noise, muscular artifacts, or other non-specific interference in EEG signals.

$$H = -\sum_{f} p(f)\log_2(p(f)), \quad p(f) = \frac{P(f)}{\sum_r P(r)} \tag{4}$$

where $P(f)$ represents power spectral density at frequency f, and $p(f) = P(f)/\sum_r P(r)$ is the relative power at f, i.e. the proportion of total spectral power contributed by that frequency.

γ-band Power: γ-band power (above 30 Hz) [41] is particularly sensitive to muscle artifacts and was quantified by summing the PSD within the γ frequency band: $\gamma_{PSD} = \sum_{f=30}^{40} P(f)$. In order to identify trials with abnormal γ-band power, a robust threshold was defined as the median of γ-band power augmented by three times its MAD. This robust measure minimises the influence of extreme outliers, thereby providing a more stable criterion of abnormality.

FOOF Analysis ($1/f$ spectrum rule): To evaluate the quality of EEG signals based on their aperiodic spectral properties, we applied the FOOF model, a spectral parameterisation method that

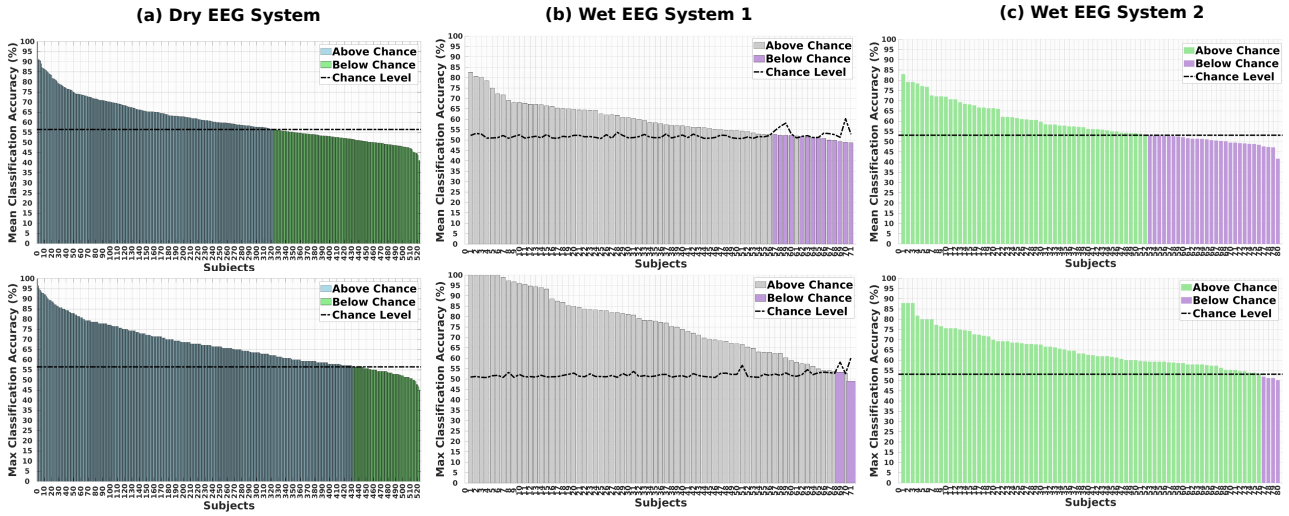

Figure 4. Sorted mean (top row) and maximum (bottom row) classification accuracy across runs for all participants with the (a) dry EEG system (Dry Training dataset), (b) wet EEG system 1 (Wet Healthy 1 and Wet AT 1 pooled together), and (c) wet EEG system 2 (Wet Healthy 2 dataset). A dashed black line shows the subject-specific chance threshold, taking into account each participant's total number of trials.

decomposes EEG power spectra into distinct aperiodic and periodic components [42]. Specifically, FOOF conceptualizes the EEG power spectrum $S(f)$ as a combination of an aperiodic component reflecting a characteristic $1/f$-like behaviour and periodic components represented by narrowband oscillatory peaks rising above this aperiodic background: $S(f) = L(f) + G(f) + \epsilon(f)$, where the aperiodic component L(f) is modelled as: $L(f) = \text{offset} - \alpha \log(f)$, with α representing the aperiodic exponent. The periodic component $G(f)$ captures oscillatory peaks, each characterized by a center frequency, amplitude, and bandwidth, while ϵ(f) reflects residual noise. This model-driven approach enables unbiased extraction of periodic oscillations and underlying broadband $1/f$-like characteristics without relying on predefined frequency bands, thus providing objective comparisons of EEG signal quality within and between subjects. Parameters including the aperiodic component's offset and exponent were extracted from each channel's PSD computed over 1-second EEG segments, facilitating the identification of substantial broadband artifacts or noise contamination. All FOOF fits were performed on each channel's 0.5–30 Hz PSD using the Python `fooof` package (v1.x). We constrained the oscillatory peaks to have widths between 1 and 12 Hz, required a minimum peak height of 0.1 μV^2/Hz, and used a peak-detection threshold of 2 standard deviations above the aperiodic fit. We also limited the maximum number of peaks to six and fixed the aperiodic mode to "fixed." Fit quality was assessed via the R^2 reported by FOOF, and any fits with $R^2 < 0.90$ were visually inspected (none were excluded).

ICLabel Artifact Contribution: In addition to the aforementioned time- and frequency-domain metrics, artifact contamination at the channel level was assessed by computing the mean artifact contribution derived from independent component (IC) analysis. Following EEG preprocessing and ICA decomposition using the extended infomax ICA algorithm [43], each IC was automatically classified using the ICLabel algorithm [44] into seven categories: brain activity, muscle activity, eye movement, cardiac artifacts, line noise, channel noise, and other sources. ICs categorized as artifacts (classes 2–7) were averaged, and their scalp projection weights were aggregated across trials, runs, and subjects.

We then constructed a channel-by-subject matrix of mean artifact contributions by stacking each electrode's aggregated IC scalp-map weights (from artifact-labeled components) across trials and subjects and expressing each channel's mean weight as a percentage of the total artifact projection across the montage. The resulting channel-wise artifact maps provided quantitative estimates of the relative severity of artifacts for each EEG channel, highlighting electrodes predominantly affected by physiological and non-physiological noise. High artifact contributions indicate substantial contamination, informing subsequent decisions regarding channel inclusion or exclusion in further EEG analyses.

Finally, for each subject, we computed a 1x7 vector of mean ICLabel class probabilities (across all ICs and trials) and then averaged these subject-level vectors to produce a concise group-level matrix reflecting the relative prevalence of brain, muscle, eye, cardiac, line noise, and other components. All channel-wise artifact maps and class-

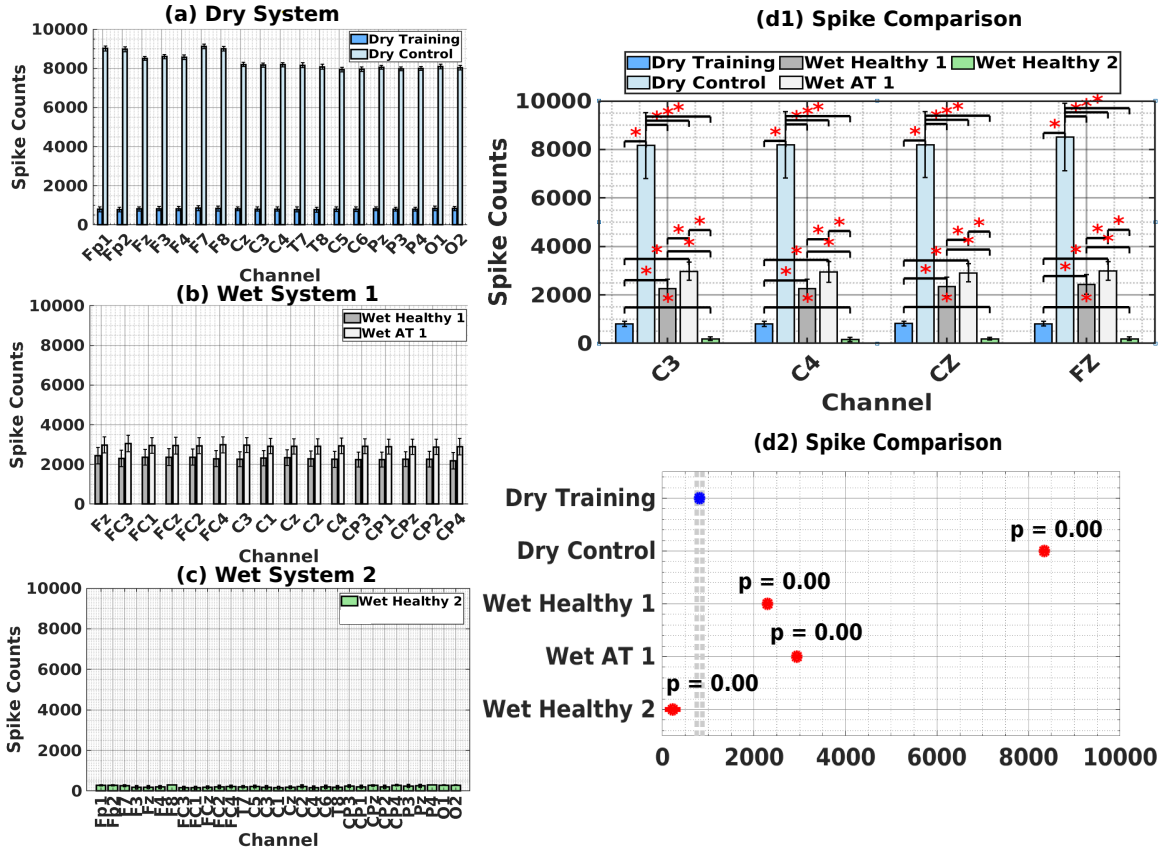

Figure 5. Spike count results across all five datasets (Dry Training, Dry Control, Wet Healthy 1, Wet AT 1, and Wet Healthy 2). (a)–(c) Channel-wise average for each EEG system as shown in the panel titles. (d1) Comparison on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red colour of the Dry Control, Wet Healthy 1, Wet AT 1 and Wet Healthy 2 distributions indicates a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

distribution matrices were obtained directly from the standard EEGLAB and ICLabel pipeline using built-in MATLAB functions [44].

2.6. Statistical Comparison between Dry and Wet EEG Systems

To isolate electrode-technology effects eliminating the confounding factor of different channel layouts across datasets, we performed a common-channel analysis restricted to C3, C4, Cz, Fz present across all systems. To objectively evaluate differences in EEG signal quality between dry and wet electrode systems, we then applied a universal testing framework with explicit effect-size reporting and corrections for multiple comparisons: first, on these four common channels, we compared the dry and wet datasets for each metric using Mann–Whitney U. Given the independent subject groups and the fact that the distributions of metrics are not always normal, this non-parametric test is appropriate. We report exact p-values after Bonferroni correction (adjusted significance threshold $\alpha_{\text{corr}} = 0.05/N_{\text{tests}}$) and accompany each comparison with the rank-biserial correlation (r_{rb}) as an effect size, facilitating interpretation of practical significance.

Second, to measure overall group-level differences per metric, we conducted one-way ANOVAs on each quality metric. The output variable instances are the corresponding metric values of each participant, averaged over channels and (data windows or trials) within subject runs in each dataset. Significant ANOVAs ($p < 0.05$) were followed by Bonferroni-corrected post-hoc (using Tukey–Kramer HSD with $\alpha = 0.05$) pairwise comparisons. For each pairwise comparison, we report the adjusted p-value and Cohen's d as an effect size, offering a clear sense of the magnitude of differences.

Together, this approach—exact, corrected p-values; explicit rank-biserial and Cohen's d effect sizes—ensures a transparent, statistically rigorous

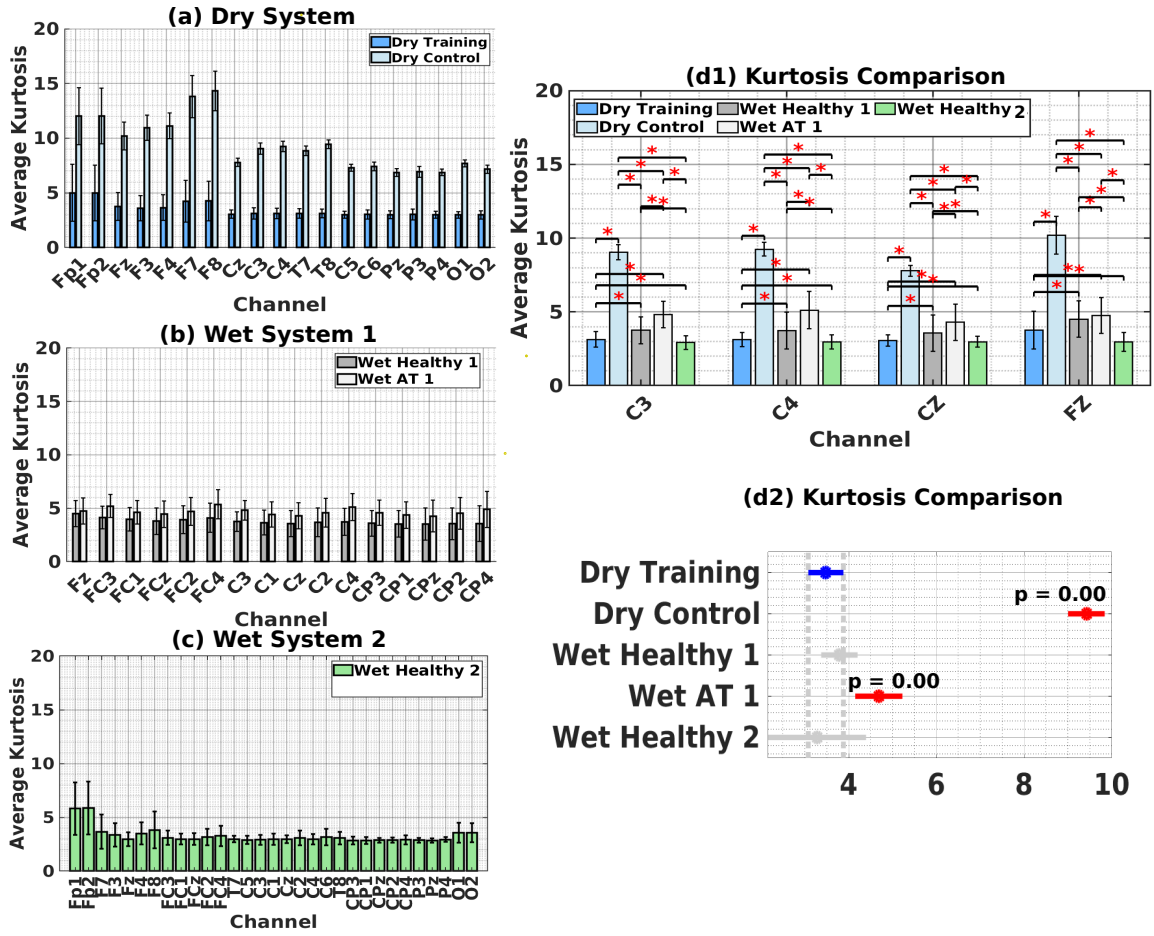

Figure 6. Kurtosis results across all five datasets. (a)–(c) Channel-wise average for each EEG system as shown in the panel titles. (d1) Comparison on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red colour of the Dry Control and Wet AT 1 distributions indicates a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Distributions shown in grey did not differ significantly from Dry Training in the post-hoc comparisons. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

comparison of dry and wet EEG systems, emphasizing both statistical validity and real-world relevance.

3. Results

Classification Accuracy: Fig. 4 contrasts the open-loop (offline), cross-validated classification accuracy achieved by participants wearing the dry-electrode EEG system (panels a) and those using the wet systems (panels b and c). When using the dry EEG system (left panels), a substantial subset of participants (324 out of 530, 61%) recorded mean accuracies above this random classification threshold (56.5% uniformly across participants for Dry Training as all subjects have the same amount of data), with several individuals reaching or surpassing 80–90%. Nonetheless, a noticeable proportion of users (39%) still fell below chance, confirming that a large portion of the population may have struggled with MI control [29] using the dry cap under the stressful and noisy conditions in which these data were acquired. Observing the maximum accuracy in the lower panel further reveals that a solid majority of participants (431 out of 570, 81%) were able to reach high performance in at least one of their runs, implying that with additional user training or more refined signal processing and machine learning, the overall classification outcome could improve for at least some part of this cohort.

By comparison, the wet EEG systems (wet EEG system 1 is the g.USBamp comprising Wet Healthy 1 and Wet AT 1 datasets, and wet EEG system 2 is the BrainAmp DC with dataset Wet Healthy 2) also show a wide range of performance levels.

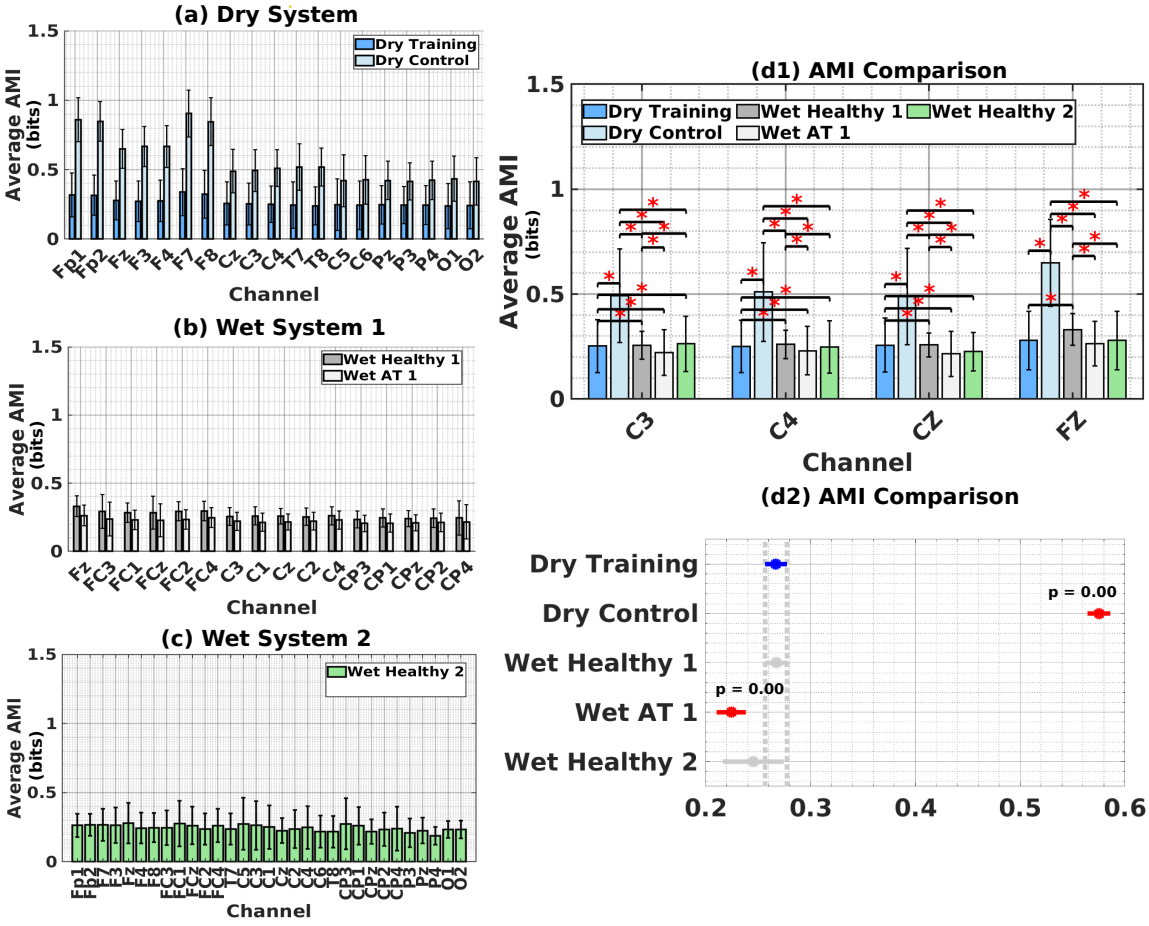

Figure 7. AMI results across all five datasets. (a)–(c) Channel-wise average for each EEG system as shown in the panel titles. (d1) Comparison on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red colour of the Dry Control and Wet AT 1 distributions indicates a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Distributions shown in grey did not differ significantly from Dry Training in the post-hoc comparisons. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

Specifically, 57 out of 71 users (80%) and 52 out of 80 users (65%) exceeded the chance level threshold for their mean accuracies, respectively, while the corresponding maximum accuracies tend to be higher at (69/71, 97% and 76/80, 95%). Because the wet-electrode recordings were generally acquired in a more controlled environment, the distribution of accuracies is somewhat narrower—few participants exhibit extremely low values. However, the shape of the curves is broadly similar to the dry-electrode case, indicating a comparable spread of individual differences in BCI aptitude. More formally, by means of a chi-squared statistical test for proportions, it is found that the difference between the dry system and the wet EEG system 1 in terms of the average classification accuracy (61% vs. 80%) is significant ($p = 0.0015$). The difference is also significant ($p = 0.0006$) in terms of maximum accuracy (81% vs. 97%). When compared to Wet EEG system 2, no significant difference is found ($p = 0.4871$) for the average accuracy (61% vs. 65%), but a significant one ($p = 0.0019$) is noted for the maximum classification accuracy (81% vs. 95%).

Collectively, our results demonstrate that mean and peak accuracies achieved with the dry systems are broadly comparable to those obtained with wet electrodes, notwithstanding the fact that the latter seem to maintain a competitive edge. To place these findings in context, we further attempt a comparison with earlier reports of MI-BCI performance collected with wet sensors under exhibition-style conditions [45]. As reported thereby, a 40-trial open-loop MI training (yielding a 70% chance-level threshold according to Müller et al. [30]) resulted in 58.6% of the open-

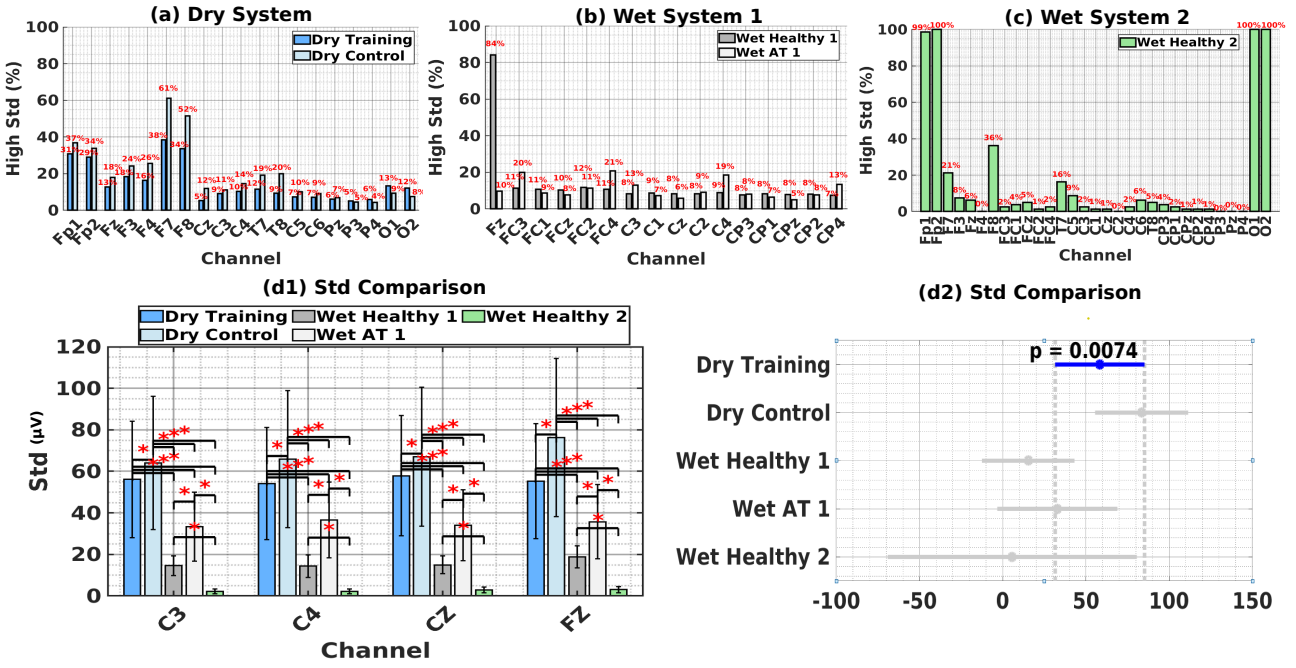

Figure 8. Standard deviation (SD) results across all five datasets. (a)–(c) Channel-wise average of the percentage of 1-second windows with abnormal standard deviation for each EEG system as shown in the panel titles. (d1) Comparison of standard deviation values on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset-dependent EEG standard deviation distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red-coloured distributions indicate a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Distributions shown in grey did not differ significantly from Dry Training in the post-hoc comparisons. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

loop sessions by 99 healthy participants exceeding the chance-level threshold with a recursive least squares algorithm and a similar 62.5% with a band-power estimation algorithm. Our results with a dry cap under substantially harsher conditions achieved a practically indistinguishable 61% figure on the same criterion out of the 530 users/sessions included, with a comparable amount of 60 trials per participant. This comparison supports our claims of an emerging competitiveness of dry EEG as opposed to traditional, wet sensors.

Spiking Activity:

Dry Control exhibits a vastly elevated spike profile across frontal, central, and temporal electrodes, whereas Dry Training is lower and more uniform; Wet Healthy 2 exhibits the lowest spike counts, followed by Dry Training, with Wet Healthy 1 and Wet AT 1 higher (Fig. 5(a)-(c)). On the four common channels (C3, C4, Cz, Fz; Fig. 5d1), all pairwise Mann–Whitney U tests were significant after Bonferroni correction ($p_{\text{adj}} < 0.0001$), with rank–biserial correlations near unity (Table 2). Fig. 5(d2) shows that group means differ (one-way ANOVA on subject-level averages: $F(4, 1841) = 7909.732$, $p < 0.0001$), and post-hoc comparisons indicate significant differences among all groups.

Kurtosis: Dry Control shows particularly high kurtosis in several channels—most prominently at frontal, central, and temporal electrodes. In Dry Training, kurtosis is relatively moderate, peaking slightly at frontal sites. The three wet datasets generally exhibit kurtosis values comparable to those of the Dry Training dataset (Fig. 6(a)–(c)). All corrected Mann–Whitney U tests on the four common channels (Fig. 6(d1)) were highly significant ($p_{\text{adj}} < 0.0001$), with rank-biserial correlations ranging from moderate ($|r_{rb}| \approx 0.53$) to near-perfect ($|r_{rb}| \geq 0.90$), reflecting substantial group differences (Table 2) both in the statistical and effect size sense. The one-way ANOVA analysis (Fig. 6(d2)) reveals a significant difference ($F(4, 1841) = 130.028$, $p < 0.0001$) across groups. Subsequent post-hoc pairwise tests further clarify which datasets diverge (mainly Dry Control and Wet AT 1), underscoring that these five recording conditions differ meaningfully in their overall kurtosis.

AMI: According to Fig. 7(a)–(c), Dry Control's AMI appears relatively low to moderate across many channels except frontal channels. By contrast, Dry

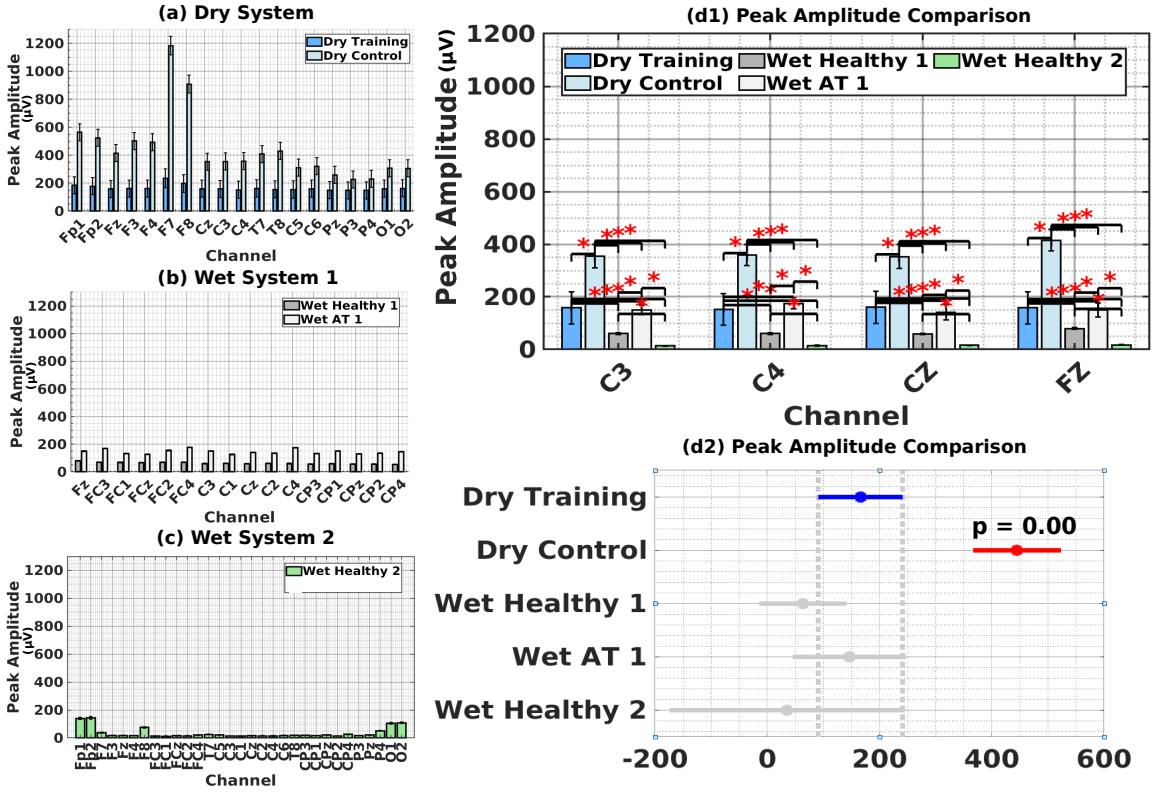

Figure 9. Peak amplitude results across all five datasets. (a)–(c) Channel-wise average for each EEG system as shown in the panel titles. (d1) Comparison on the set of channels common to all datasets, with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red-coloured distributions indicate a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Distributions shown in grey did not differ significantly from Dry Training in the post-hoc comparisons. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

Training, Wet AT 1, Wet Healthy 1, and Wet Healthy 2 exhibit low AMI values across all the channels. Fig. 7(d1), depicting the four common channel comparisons, aligns with the findings on the previous metrics, showing considerably and statistically significant greater AMI for Dry Control especially in Fz ($p_{\text{adj}} < 0.0001$, with rank–biserial correlations from $|r_{rb}|\approx 0.20$ to ≥ 0.90—except for the comparison between Wet AT 1 and Wet Healthy 2, see Table 2). The remaining four datasets are closer to each other in this case. Turning to the one-way ANOVA (Fig. 7(d2)), each participant's mean AMI (averaged over their channels) differs significantly ($F(4, 1841) = 611.095$, $p < 0.0001$) across the five conditions.

Standard Deviation: In the Dry System, Dry Control features a noticeably higher percentage of channels flagged as abnormal, particularly in frontal and temporal electrodes, relative to Dry Training. Wet Healthy 1 and Wet AT 1 differ mainly at frontal–central sites, with higher abnormal rates for the healthy group and Wet Healthy 2 shows moderate abnormal percentages at frontal and occipital sites, generally lower than those observed in Dry Control and Wet Healthy 1 (Fig. 8(a)–(c)). In the four common channels (Fig. 8d1), although Dry Control remains the condition most influenced by high-amplitude noise, Dry Training also shows significantly higher standard deviation values, especially in frontal–central channels, exceeding those of Wet AT 1, Wet Healthy 1, and Wet Healthy 2. All Mann–Whitney U tests (Bonferroni-corrected) on standard deviation at C3, C4, Cz and Fz were highly significant ($p_{\text{adj}} < 0.0001$), with rank–biserial correlations from moderate ($|r_{rb}| \approx 0.27$) to maximal ($|r_{rb}| = 1.00$) (Table 2). Turning to one-way ANOVA (Fig. 8d2), each participant's mean standard deviation (averaged across all channels) differs significantly ($F(4, 1841) = 3.503$, $p = 0.0074$) among the five datasets. However, any dissimilarities of the underlying distributions seem to be marginal, as post-hoc comparisons do not show a significant

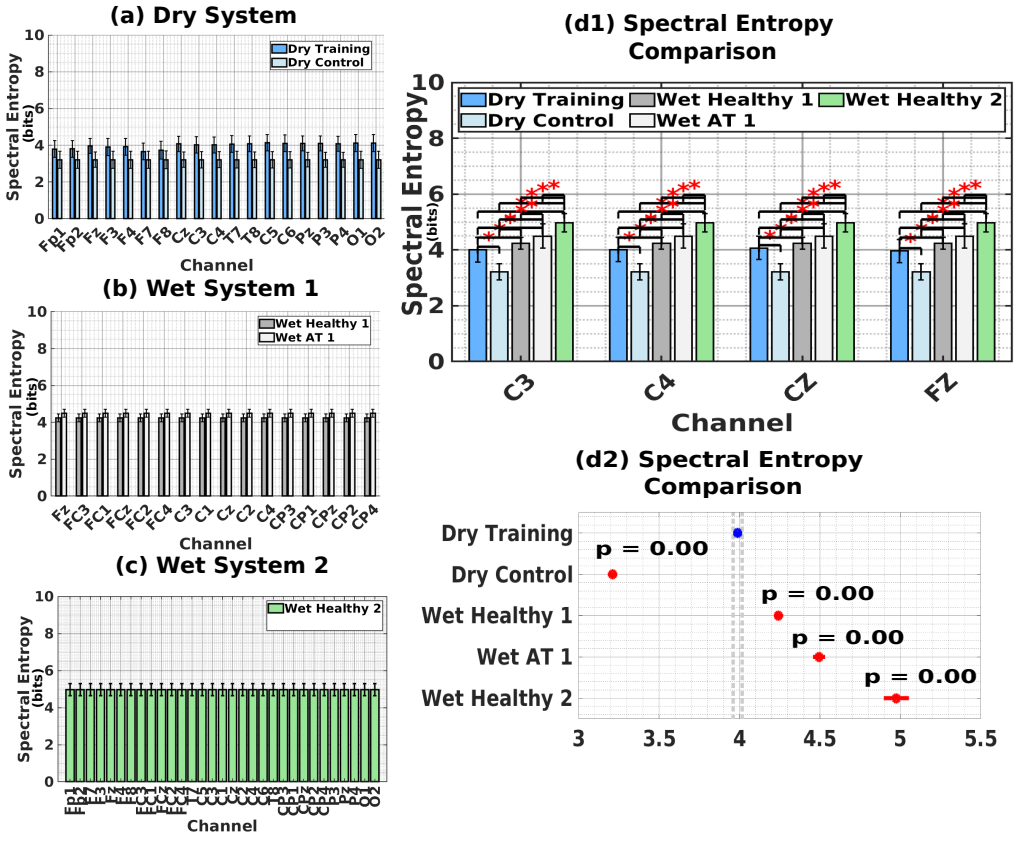

Figure 10. Spectral Entropy results across all five datasets. (a)–(c) Channel-wise average for each EEG system as shown in the panel titles. (d1) Comparison on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red-coloured distributions indicate a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

difference for any specific group pair.

Peak Amplitude: Several channels—particularly frontal (F7, F8) and temporal—exhibit notably large peak amplitudes in Dry Control. By contrast, Dry Training shows comparatively moderate values across most channels and is closely aligned with Wet AT 1. Both Wet Healthy 1 and Wet Healthy 2 display low peaks across the majority of channels (Fig. 9(a)–(c)). For four common channels, pairwise Mann–Whitney U tests among all five datasets were significant after Bonferroni correction ($p_{\text{adj}} < 0.0001$), with rank–biserial correlations from moderate to near-perfect ($|r_{rb}| = 0.54$–1.00; Table 2). Mean peak amplitudes differed significantly among datasets (one-way ANOVA: $F(4, 1841) = 13.367$, $p < 0.0001$), and post-hoc comparisons clarified which pairs diverged most (e.g., *Dry Training* vs. *Dry Control*: $p_{\text{adj}} = 0.0000$, Cohen's $d = -0.24$) (Table 4).

Spectral Entropy:

Dry Control recordings exhibit markedly smaller spectral entropy, with values generally around or below 3.5 bits across most channels. The Dry Training dataset shows moderate spectral entropy relative to Dry Control. In contrast, all wet datasets (Wet Healthy 1, Wet AT 1, Wet Healthy 2) exhibit higher spectral entropy, consistently approaching or exceeding values near 4.3–5 (Fig. 10(a)–(c)). For C3, C4, Cz, Fz (Fig. 10d1), pairwise Mann–Whitney U comparisons reveal consistent significant differences (Bonferroni-corrected $p_{\text{adj}} < 0.0001$) across multiple dataset pairs (Table 3). A one-way ANOVA on subject-level mean spectral entropy (averaged across channels) confirms group differences ($F(4, 1841) = 1094.904$, $p < 0.0001$; Fig. 10d2).

γ-band power: Fig. 11(a)–(c) illustrate threshold-based comparisons of γ-band power—namely, the percentage of participants in each dataset whose γ-band PSD at a given channel exceeds a robust threshold. Dry Control shows remarkably higher

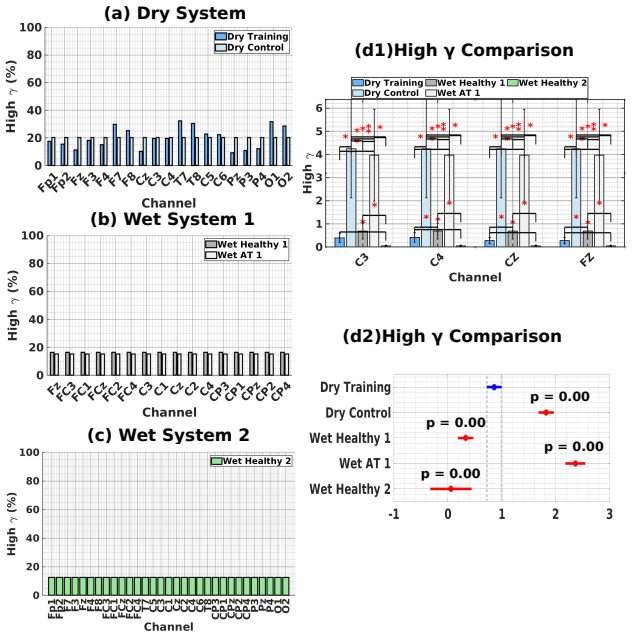

Figure 11. γ-band power results across all five datasets. (a)–(c) Channel-wise average of time windows with abnormal γ-band power for each EEG system as shown in the panel titles. (d1) Comparison of γ-band power on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset-dependent γ-band power distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red-coloured distributions indicate a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

γ-band components, especially at peripheral sites (frontal, temporal, occipital). In Wet Healthy 1 and Wet AT 1, abnormal γ appears in less than 20% of the data across all channels; Wet Healthy 2 exhibits even lower percentages, with small numerical differences. On the four common channels (Fig. 11d1), all dataset-based differences are statistically significant ($p_{\text{adj}} < 0.0001$), with rank–biserial correlations indicating very strong separations (Table 3); in this view, Wet AT 1 is more affected than *Wet Healthy 1*. A one-way ANOVA on each participant's mean γ-band power (averaged across channels) indicates group differences ($F(4, 1841) = 105.767$, $p < 0.0001$; Fig. 11d2). Post-hoc tests show that all groups differ significantly from each other, with larger effect sizes for Dry Control and Wet AT 1.

FOOF Analysis: The aperiodic component of the spectrum of EEG is known to be modulated both by various mental tasks [42] and linked to the signal's quality. Exactly how the signal integrity influences the spectrum shape is not entirely understood, however, empirical evidence [42] suggests that noise on EEG may manifest either as a flatter spectrum (i.e., larger aperiodic exponent as it appears in the FOOF model) and potentially also larger offset (i.e., a negative value closer to 0 given the logarithmic representation in FOOF). Our results are consistent with this, as Dry Control and Training caused on average a flattening of the spectrum's shape as shown by exponent values in the range 1.5-2.0 or above, and concomitant elevation of the offset relative to the wet electrodes (Wet Healthy 1 and 2, and Wet AT 1), notwithstanding the fact that all these differences are not so pronounced. The second wet configuration (Wet Healthy 2) is somewhat peculiar as it exhibited flatter spectra for frontal channels, but, at the same time, the smallest offset across all conditions (Fig. 12(a)–(c)). The detailed four-channel comparison (Fig. 12(d1)) verifies the same effects and highlights that differences are statistically significant (Mann–Whitney U, Bonferroni-corrected, $p_{\text{adj}} < 0.0001$) across several pairwise dataset combinations (Table 3). Of note, the aperiodic exponent results are in agreement with those on spectral entropy, corroborating the sanity of this analysis, as the two metrics convey the same type of information. Fig. 12(d2) summarizes the one-way ANOVA analyses across all five datasets, revealing significant differences in overall group means for both FOOF aperiodic exponent ($F(4, 1841) = 673.852$, $p < 0.0001$) and offset ($F(4, 1841) = 2000.303$, $p < 0.0001$).

ICLabel Artifact Contribution: The ICLabel decomposition analysis revealed distinct, condition-specific artifact patterns across EEG systems (Fig. 13). In the Dry Training dataset (Fig. 13(a)), a moderate artifact influence (maximum $\approx 1\%$) was most pronounced over frontal (Fp1, Fp2) and occipital-temporal (O2, T8) electrodes, while central electrodes (Cz, C3, C4) exhibited relatively lower artifact contributions. Class-wise, ICLabel attributed approximately 31% of retained variance to brain activity, while the second largest proportion of artifacts originated from the "Other" category (29.5%), with moderate contributions from muscle (10.6%), line noise (13.5%), and ocular (6.2%) sources. During the noisier Dry Control condition (Fig. 13(b)), artifact magnitude significantly increased, reaching above 1.5% prominently at bilateral temporal sites (T7, T8, C5, C6) and frontal regions. Nevertheless, the overall proportion of brain-related ICs remained practically the same as for Dry Training (33.5%), as also muscular artifacts did (9.2%). Overall, it is shown that, except for peripheral channels that are more considerably affected in Dry Control, the

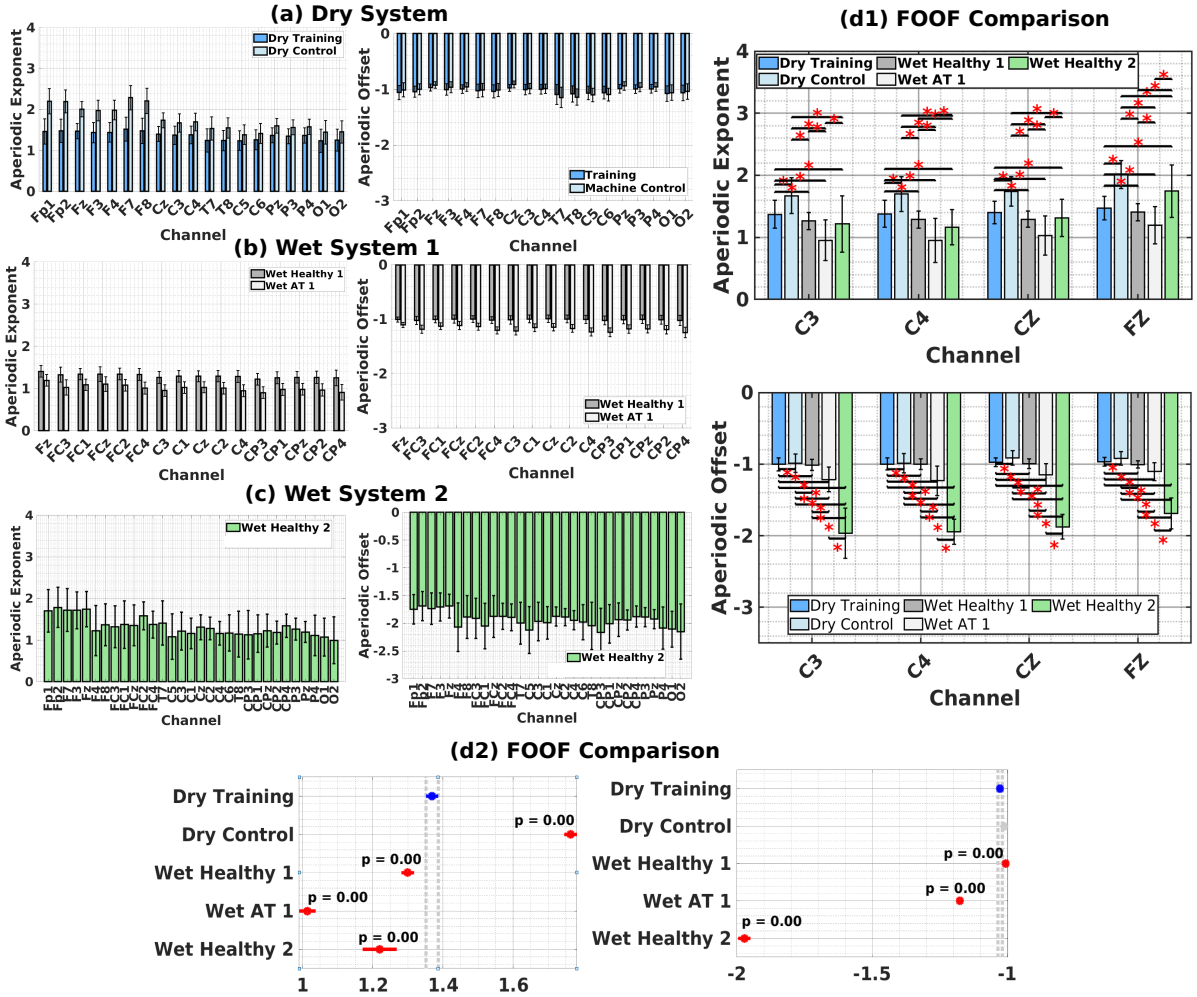

Figure 12. FOOF-derived offset and exponent of the aperiodic component of EEG across all five datasets. (a)–(c) Channel-wise average for each EEG system as shown in the panel titles. (d1) Comparison on the set of channels common to all datasets with red asterisks indicating significant differences based on Mann–Whitney U tests (Bonferroni corrected). (d2) Visualisation of dataset distributions (mean and standard deviation). Blue colour of the Dry Training distribution indicates significance ($p < 0.05$) of the one-way ANOVA effect with factor "dataset". Red-coloured distributions indicate a significant difference of the Tukey-Kramer post-hoc test comparing the respective distribution with that of Dry Training. Distributions shown in grey did not differ significantly from Dry Training in the post-hoc comparisons. Cohen's d effect sizes for panel (d2)—each dataset vs. Dry Training, with asterisks marking Tukey–Kramer–adjusted $p < 0.05$—are summarized in Table 4 (sign indicates direction relative to Dry Training).

influence of noise remains largely similar to Dry Training. The only pronounced difference is a redistribution of noise type in favour of the "Other" category (39% vs 29.5%), which should reflect the anticipated greater scale of movement- and stress-related artifacts in this condition, consistent with public, closed-loop use of devices. In contrast, the controlled laboratory Wet Healthy 1 dataset presented the cleanest artifact profile, with contributions generally below 3%, distributed relatively uniformly across central and frontal areas. ICLabel identified a high proportion of brain components (49.5%) and lower percentages of muscle (8%), ocular (4.5%), and channel noise (2.3%) sources (Fig. 13 (c)). Finally, the semi-controlled Wet AT 1 recordings showed a distinct posterior-temporal artifact distribution, with peak artifact levels around 1% at parietal sites (CP4, CP2). The overall brain fraction decreased to 41%, while eye-related artifacts notably increased (5.7%), muscle activity was moderate (10.5%), and a higher percentage (22.8%) was classified as "Other", suggesting mixed artifact sources typical of clinical or home settings (Fig. 13 (d)). Finally, in the high-density, fully controlled Wet Healthy 2 dataset (Fig. 13(e)), artifact contributions were again quite low overall (<1%), but showed a somewhat broader spatial footprint than Wet Healthy 1, with modest peaks at fronto-central (FC3/FC4) and parieto-occipital (P3/P4, O1/O2) sites. ICLabel attributed 35.3% of the variance to brain sources, with the largest single artifact category now being "Other". Muscle and

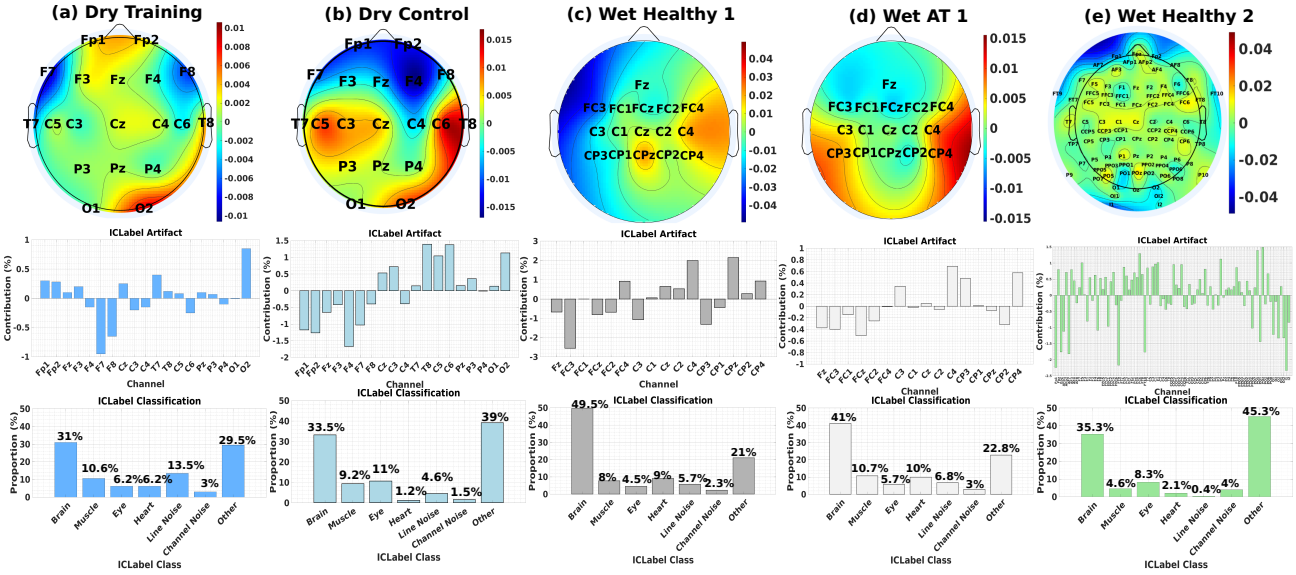

Figure 13. ICLabel summary across five analyzed EEG datasets. Top row: scalp topographies illustrating the spatial distribution of artifact contributions across EEG channels. Dark red colors indicate higher artifact presence, while dark blue colours represent lower artifact influence. Middle row: bar plots quantifying per-channel artifact contributions (%). Bottom row: bar charts representing the overall percentage of retained variance classified into seven ICLabel categories (Brain, Muscle, Eye, Heart, Line Noise, Channel Noise, Other). Columns correspond to (a) Dry Training, (b) Dry Control, (c) Wet Healthy 1, (d) Wet AT 1 and Wet Healthy 2 datasets.

channel-noise contributions were both under 5%, eye artifacts rose to 8.3%, heart activity remained minimal, and line-noise contributions were negligible. Overall, the dry headset yielded component mixtures comparable to wet recordings, but was more susceptible to channel noise (during training) and to movement artifacts when used in the acoustically and electromagnetically challenging exhibition environment.

4. Discussion

While several studies [7, 2, 4, 5, 6, 9, 1] have begun to close the literature gap in benchmarking dry and wet electrodes, they remain limited to small cohorts (N<100), mostly ERP/SSVEPs paradigms, and fixed lab setups. By contrast, this work contributes a MI BCI comparison in a realistic and very challenging (as far as noise and physiological artifact sources are concerned) public exhibition setting, with half a thousand users. We not only show that dry caps can perform on par with wet systems, but we also map out the residual artifact landscape that emerges when BCIs are taken out of the lab—a critical step toward truly "in-the-wild" neurotechnology.

4.1. Classification Performance: Dry vs. Wet

The classification accuracy results underscore that, although wet EEG traditionally provides a high signal-to-noise ratio in controlled settings, state-of-the-art dry electrodes can deliver comparable performance. Both systems show a portion of individuals who readily achieve robust MI-based control, while others remain below the threshold. The distinction is that the dry-electrode recordings were obtained under realistic—and often noisy—conditions, which might have penalized average performance relative to more standardized laboratory protocols used in the majority of the wet-electrode dataset recordings. Although the ensemble of these results still suggests a marginal superiority of wet electrodes, the fact that a substantial subset of the dry-system users 61% attained above-chance classification, and that their maximum accuracies approached the upper range of the wet-system participants, supports the fact that the latest dry technology is well-suited for BCI usage outside traditional research facilities. This aligns with earlier observations that user-to-user performance variability is nowadays a more serious concern than the choice of electrode type, especially when artifact detection and user training are in place.

4.2. Comprehensive Signal–Quality Assessment

We further proceeded with a comprehensive assessment across multiple EEG signal quality metrics, revealing the critical roles of the recording environment, electrode technology, and task-specific conditions in shaping EEG signal integrity.

The markedly elevated spike profile in Dry Control is most definitely reflecting the noisier conditions and

Table 2. Mann–Whitney U tests (Bonferroni corrected) for time-domain metrics.

Metric	Comparison	C3			C4			Cz			Fz		
		p_{corr}	U	r_{rb}	p_{corr}	U	r_{rb}	p_{corr}	U	r_{rb}	p_{corr}	U	r_{rb}
Spike Count													
	Dry Training vs. Dry Control	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00
	Dry Training vs. Wet AT 1	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00	0.00	0	1.00
	Dry Training vs. Wet Healthy 1	0.00	242	1.00	0.00	953	0.99	0.00	390	1.00	0.00	241	1.00
	Dry Training vs. Wet Healthy 2	0.00	42 332	−1.00	0.00	42 356	−1.00	0.00	42 334	−1.00	0.00	42 325	−1.00
	Dry Control vs. Wet AT 1	0.00	131 175	−1.00	0.00	131 175	−1.00	0.00	131 175	−1.00	0.00	131 175	−1.00
	Dry Control vs. Wet Healthy 1	0.00	230 868	−1.00	0.00	230 868	−1.00	0.00	230 868	−1.00	0.00	230 868	−1.00
	Dry Control vs. Wet Healthy 2	0.00	38 160	−1.00	0.00	38 160	−1.00	0.00	38 160	−1.00	0.00	38 160	−1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	22 000	−1.00	0.00	22 000	−1.00	0.00	22 000	−1.00	0.00	22 000	−1.00
	Wet Healthy 1 vs. Wet Healthy 2	0.00	38 720	−1.00	0.00	38 720	−1.00	0.00	38 720	−1.00	0.00	38 720	−1.00
Kurtosis													
	Dry Training vs. Dry Control	0.00	2 848	0.98	0.00	1 989	0.98	0.00	1 939	0.98	0.00	8 648	0.93
	Dry Training vs. Wet AT 1	0.00	12 981	0.82	0.00	11 890	0.84	0.00	12 519	0.83	0.00	34 117	0.53
	Dry Training vs. Wet Healthy 1	0.00	26 161	0.80	0.00	32 868	0.74	0.00	30 136	0.77	0.00	56 664	0.56
	Dry Training vs. Wet Healthy 2	0.00	30 544	−0.44	0.00	29 256	−0.38	0.00	26 697	−0.26	0.00	34 507	−0.63
	Dry Control vs. Wet AT 1	0.00	115 177	−0.76	0.00	111 647	−0.70	0.00	118 085	−0.80	0.00	124 454	−0.90
	Dry Control vs. Wet Healthy 1	0.00	230 868	−1.00	0.00	230 868	−1.00	0.00	230 868	−1.00	0.00	230 868	−1.00
	Dry Control vs. Wet Healthy 2	0.00	38 160	−1.00	0.00	38 160	−1.00	0.00	38 160	−1.00	0.00	38 160	−1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	20 645	−0.88	0.00	20 490	−0.86	0.00	20 220	−0.84	0.00	20 778	−0.89
	Wet Healthy 1 vs. Wet Healthy 2	0.00	36 328	−0.88	0.00	34 376	−0.78	0.00	34 599	−0.79	0.00	36 788	−0.90
AMI													
	Dry Training vs. Dry Control	0.00	25 236	0.80	0.00	23 986	0.81	0.00	23 372	0.82	0.00	9 245	0.93
	Dry Training vs. Wet AT 1	0.00	93 267	−0.28	0.00	87 200	−0.20	0.00	100 857	−0.38	1.00	75 260	−0.03
	Dry Training vs. Wet Healthy 1	0.00	82 867	0.35	0.00	77 726	0.39	0.00	82 015	0.36	0.00	54 865	0.57
	Dry Training vs. Wet Healthy 2	0.00	25 998	−0.23	0.00	27 558	−0.30	0.00	27 321	−0.29	0.02	24 603	−0.16
	Dry Control vs. Wet AT 1	0.00	120 756	−0.84	0.00	119 136	−0.82	0.00	122 276	−0.86	0.00	127 528	−0.94
	Dry Control vs. Wet Healthy 1	0.00	206 643	−0.79	0.00	206 131	−0.79	0.00	208 280	−0.80	0.00	220 259	−0.91
	Dry Control vs. Wet Healthy 2	0.00	32 430	−0.70	0.00	33 463	−0.75	0.00	34 580	−0.81	0.00	35 603	−0.87
	Wet AT 1 vs. Wet Healthy 2	0.75	10 743	0.02	0.69	11 319	−0.03	0.49	10 440	0.05	0.30	11 840	−0.08
	Wet Healthy 1 vs. Wet Healthy 2	0.00	25 808	−0.33	0.00	28 065	−0.45	0.00	26 893	−0.39	0.00	28 224	−0.46
Standard Deviation													
	Dry Training vs. Dry Control	0.00	17 969	0.86	0.00	17 393	0.86	0.00	15 886	0.87	0.00	9 913	0.92
	Dry Training vs. Wet AT 1	0.00	14 908	0.80	0.00	20 208	0.72	0.00	17 110	0.77	0.00	16 670	0.77
	Dry Training vs. Wet Healthy 1	0.00	86 252	0.33	0.00	93 859	0.27	0.00	76 511	0.40	0.00	53 991	0.58
	Dry Training vs. Wet Healthy 2	0.00	42 271	−0.99	0.00	42 048	−0.98	0.00	42 381	−1.00	0.00	42 288	−0.99
	Dry Control vs. Wet AT 1	0.00	84 911	−0.29	0.00	84 025	−0.28	0.00	90 574	−0.38	0.00	110 998	−0.69
	Dry Control vs. Wet Healthy 1	0.00	219 886	−0.90	0.00	219 351	−0.90	0.00	216 578	−0.88	0.00	222 723	−0.93
	Dry Control vs. Wet Healthy 2	0.00	38 160	−1.00	0.00	38 152	−1.00	0.00	38 160	−1.00	0.00	38 160	−1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	22 000	−1.00	0.00	21 974	−1.00	0.00	22 000	−1.00	0.00	22 000	−1.00
	Wet Healthy 1 vs. Wet Healthy 2	0.00	38 690	−1.00	0.00	38 643	−1.00	0.00	38 720	−1.00	0.00	38 694	−1.00
Peak Amplitude													
	Dry Training vs. Dry Control	0.00	6 266	0.95	0.00	5 669	0.96	0.00	7 271	0.94	0.00	4 307	0.97
	Dry Training vs. Wet AT 1	0.00	7 668	0.89	0.00	11 091	0.85	0.00	8 727	0.88	0.00	10 654	0.85
	Dry Training vs. Wet Healthy 1	0.00	58 543	0.54	0.00	57 289	0.55	0.00	44 125	0.66	0.00	38 563	0.70
	Dry Training vs. Wet Healthy 2	0.00	41 581	−0.96	0.00	41 367	−0.95	0.00	42 015	−0.98	0.00	40 582	−0.91
	Dry Control vs. Wet AT 1	0.00	99 085	−0.51	0.00	94 868	−0.45	0.00	101 142	−0.54	0.00	119 657	−0.82
	Dry Control vs. Wet Healthy 1	0.00	225 890	−0.96	0.00	225 995	−0.96	0.00	225 098	−0.95	0.00	227 888	−0.97
	Dry Control vs. Wet Healthy 2	0.00	38 160	−1.00	0.00	38 046	−0.99	0.00	38 160	−1.00	0.00	38 160	−1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	22 000	−1.00	0.00	21 829	−0.98	0.00	22 000	−1.00	0.00	21 960	−1.00
	Wet Healthy 1 vs. Wet Healthy 2	0.00	38 615	−0.99	0.00	38 109	−0.97	0.00	38 648	−1.00	0.00	38 453	−0.99

increased movements during active machine operation. The ordering among wet datasets (Wet Healthy 1 vs. Wet AT 1) is consistent with the fact that the latter were recorded with the same system as Wet Healthy 1, but in harsher (clinical or home) conditions. Within the exhibition venue, the very good performance of the dry cap during Dry Training in the same overall setting of the Mental Work exhibition venue must be attributed to the extra demands of the machine BCI control task and the additionally burdened surroundings (i.e., crowds observing in close proximity, talking loudly, carrying cell phones, etc.) during Dry Control. Taken together, the trends (with Wet Healthy 2 lowest, closely followed by Dry Training) further hint at the emerging competitiveness of dry sensors under calibration-like conditions, when environmental load is moderated. On the other hand, the elevated kurtosis at frontal, central, and temporal sites in Dry Control is compatible with episodic large-amplitude transients (e.g., muscle bursts), while Dry Training and the wet datasets show lower, more uniform values. Despite some attenuation in

Table 3. Mann–Whitney U tests (Bonferroni corrected) for frequency-domain metrics.

Metric	Comparison	C3			C4			Cz			Fz		
		p_{corr}	U	r_{rb}	p_{corr}	U	r_{rb}	p_{corr}	U	r_{rb}	p_{corr}	U	r_{rb}
Spectral Entropy													
	Dry Training vs. Dry Control	0.00	238056	−0.88	0.00	238942	−0.89	0.00	242022	−0.91	0.00	238318	−0.89
	Dry Training vs. Wet AT 1	0.00	26379	0.64	0.00	25619	0.65	0.00	28753	0.61	0.00	21307	0.71
	Dry Training vs. Wet Healthy 1	0.00	77476	0.40	0.00	75380	0.41	0.00	88017	0.31	0.00	58018	0.55
	Dry Training vs. Wet Healthy 2	0.00	945	0.96	0.00	892	0.96	0.00	1083	0.95	0.00	641	0.97
	Dry Control vs. Wet AT 1	0.00	1875	0.97	0.00	1875	0.97	0.00	1875	0.97	0.00	1875	0.97
	Dry Control vs. Wet Healthy 1	0.00	990	0.99	0.00	990	0.99	0.00	990	0.99	0.00	990	0.99
	Dry Control vs. Wet Healthy 2	0.00	2	1.00	0.00	2	1.00	0.00	2	1.00	0.00	2	1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	3834	0.65	0.00	3834	0.65	0.00	3834	0.65	0.00	3834	0.65
	Wet Healthy 1 vs. Wet Healthy 2	0.00	1638	0.92	0.00	1638	0.92	0.00	1638	0.92	0.00	1638	0.92
Gamma PSD													
	Dry Training vs. Dry Control	0.00	36462	0.71	0.00	37703	0.70	0.00	21965	0.83	0.00	22330	0.82
	Dry Training vs. Wet AT 1	0.00	13377	0.82	0.00	14 092	0.81	0.00	8394	0.88	0.00	8023	0.89
	Dry Training vs. Wet Healthy 1	0.02	138675	−0.08	0.00	145254	−0.13	0.00	98523	0.23	0.00	96854	0.24
	Dry Training vs. Wet Healthy 2	0.00	39464	−0.86	0.00	39531	−0.86	0.00	38378	−0.81	0.00	37 918	−0.79
	Dry Control vs. Wet AT 1	0.00	47285	0.28	0.00	47 285	0.28	0.00	47 285	0.28	0.00	47285	0.28
	Dry Control vs. Wet Healthy	0.00	215829	−0.87	0.00	215 829	−0.87	0.00	215829	−0.87	0.00	215829	−0.87
	Dry Control vs. Wet Healthy 2	0.00	38073	−1.00	0.00	38 073	−1.00	0.00	38073	−1.00	0.00	38073	−1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	21985	−1.00	0.00	21985	−1.00	0.00	21 985	−1.00	0.00	21985	−1.00
	Wet Healthy 1 vs. Wet Healthy 2	0.00	36709	−0.90	0.00	36709	−0.90	0.00	36709	−0.90	0.00	36709	−0.90
FOOOF Exponent													
	Dry Training vs. Dry Control	0.00	44 888	0.64	0.00	40 538	0.68	0.00	24 348	0.81	0.00	7 681	0.94
	Dry Training vs. Wet AT 1	0.00	130 721	−0.79	0.00	131 405	−0.80	0.00	131 556	−0.81	0.00	120 167	−0.65
	Dry Training vs. Wet Healthy 1	0.00	176 668	−0.38	0.00	170 551	−0.33	0.00	186 427	−0.45	0.00	160 422	−0.25
	Dry Training vs. Wet Healthy 2	0.00	25 969	−0.22	0.00	32 102	−0.51	0.01	25 871	−0.22	0.00	9 557	0.55
	Dry Control vs. Wet AT 1	0.00	126 079	−0.92	0.00	126 562	−0.93	0.00	128 439	−0.96	0.00	129 605	−0.98
	Dry Control vs. Wet Healthy 1	0.00	210 517	−0.82	0.00	211 697	−0.83	0.00	223 814	−0.94	0.00	228 724	−0.98
	Dry Control vs. Wet Healthy 2	0.00	31 627	−0.66	0.00	35 149	−0.84	0.00	33 475	−0.75	0.00	27 748	−0.45
	Wet AT 1 vs. Wet Healthy 2	0.00	5 731	0.48	0.00	6 338	0.42	0.00	4 744	0.57	0.00	2 493	0.77
	Wet Healthy 1 vs. Wet Healthy 2	0.53	18 596	0.04	0.00	25 743	−0.33	1.00	18 143	0.06	0.00	6 932	0.64
FOOOF Offset													
	Dry Training vs. Dry Control	0.00	106 696	0.16	0.01	111 827	0.12	0.00	63 786	0.50	0.00	77 266	0.39
	Dry Training vs. Wet AT 1	0.00	131 411	−0.80	0.00	129 648	−0.78	0.00	130 992	−0.80	0.00	129 011	−0.77
	Dry Training vs. Wet Healthy 1	0.00	152 402	−0.19	0.00	148 974	−0.16	0.00	164 265	−0.28	0.00	194 876	−0.52
	Dry Training vs. Wet Healthy 2	0.00	42 374	−1.00	0.00	42 396	−1.00	0.00	42 400	−1.00	0.00	42 393	−1.00
	Dry Control vs. Wet AT 1	0.00	115 130	−0.76	0.00	113 260	−0.73	0.00	121 716	−0.86	0.00	118 999	−0.81
	Dry Control vs. Wet Healthy 1	0.00	144 481	−0.25	0.00	135 367	−0.17	0.00	186 267	−0.61	0.00	187 267	−0.62
	Dry Control vs. Wet Healthy 2	0.00	38 136	−1.00	0.00	38 151	−1.00	0.00	38 158	−1.00	0.00	38 152	−1.00
	Wet AT 1 vs. Wet Healthy 2	0.00	21 766	−0.98	0.00	21 765	−0.98	0.00	21 943	−0.99	0.00	21 821	−0.98
	Wet Healthy 1 vs. Wet Healthy 2	0.00	38 694	−1.00	0.00	38 711	−1.00	0.00	38 718	−1.00	0.00	38 714	−1.00

Table 4. Cohen's d for pairwise comparisons *vs. Dry Training*. Asterisks (*) denote significant Tukey–Kramer post-hoc differences (adjusted $p < 0.05$).

Comparison (vs. Dry Training)	Spike	Kurtosis	AMI	SD	PeakAmp	SpEntropy	γPSD	FOOOF Exp/Offset
Dry Control	-8.15^{*}	-1.00^{*}	-2.10^{*}	−0.06	-0.24^{*}	2.29^{*}	0.08	-0.54^{*}
Wet Healthy 1	-5.39^{*}	−0.31	−0.00	0.11	0.10	-0.82^{*}	0.09	-0.25^{*}
Wet AT 1	-9.12^{*}	-0.50^{*}	0.37^{*}	0.06	0.02	-1.27^{*}	0.08	-0.30^{*}
Wet Healthy 2	6.71^{*}	0.29	0.18	0.10	0.10	-2.64^{*}	0.07	0.29

magnitude, the kurtosis ordering mirrors the spike-count pattern, suggesting similar context sensitivity. Furthermore, the Dry Control has relatively higher AMI at frontal sites, especially Fz, which is consistent with stronger temporal dependencies in the presence of environmental load and muscle/interference sources. Conversely, the uniformly low AMI observed for Dry Training, Wet AT 1, Wet Healthy 1, and Wet Healthy 2 reflects more stable, less autocorrelated activity. The higher standard-deviation abnormality rates observed

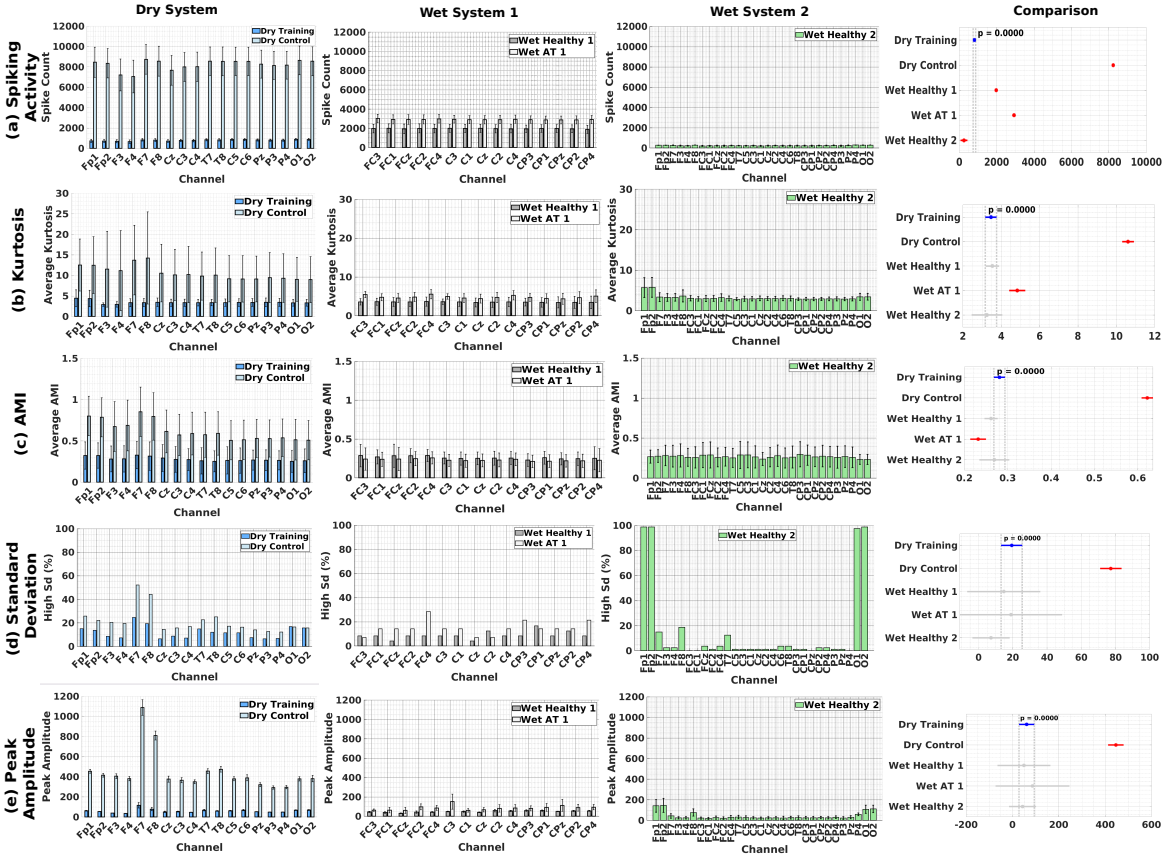

Figure 14. Different time-domain quality matrices after re-referencing to Fz for all five datasets. Panels show (a) spike count, (b) kurtosis, (c) average AMI, (d) standard deviation, and (e) peak amplitude. Despite the known artifact sensitivity of Fz, the pre- and post-reference matrices are virtually identical, with only slight variance in the standard-deviation panel, demonstrating that shape- and distribution-based metrics are insensitive to reference selection.

for Dry Control, particularly at frontal and temporal electrodes relative to Dry Training, suggest that active machine operation and a noisy public environment lead to larger amplitude variability. Compared to transient metrics (i.e., spikes), the standard-deviation metric may indicate a more persistent quality degradation, consistent with the peripheral channels repeatedly exhibiting high abnormal rates. Notably, standard deviation is the first metric where both types of dry data Dry Control and Dry Training) collectively perform less favourably—albeit by a small margin—than the wet datasets, suggesting that electrode technology may play a relatively larger role here than environmental noise alone. For peak amplitude, large peaks at frontal and temporal sites in Dry Control suggest abrupt bursts of signal, possibly tied to movement or muscle artifacts in an active public environment. The similarity between Dry Training and Wet AT 1 indicates that, for this metric, the environmental factor can be more critical than the electrode technology alone. Higher spectral entropy reflects spectra that follow the canonical $1/f$ structure with expected neural peaks (e.g., α band), whereas lower values are consistent with flatter, artifact-laden spectra. In this light, the markedly lower entropy in Dry Control aligns with heavier contamination, while the wet datasets show consistently higher entropy. Notably, Dry Training approaches Wet Healthy 1, indicating that under a moderated environmental load, the gap narrows. In contrast to spectral entropy, the γ-band power metric seems heavily dependent on the environmental conditions rather than the EEG technology.

Notably, the spectral entropy analysis highlighted lower entropy values in the Dry Control dataset, indicative of lower signal quality, whereas wet-electrode data displayed broader spectral complexity, typical of richer neural information captured in controlled environments. The findings from the FOOF analysis further corroborate these observations, demonstrating less favourable aperiodic exponent and offset values in dry systems during active task engagement (Dry Control) and clinical conditions (Wet AT 1), underscoring a greater impact of environmental influences compared to the inherent technological shortcomings. The differential patterns

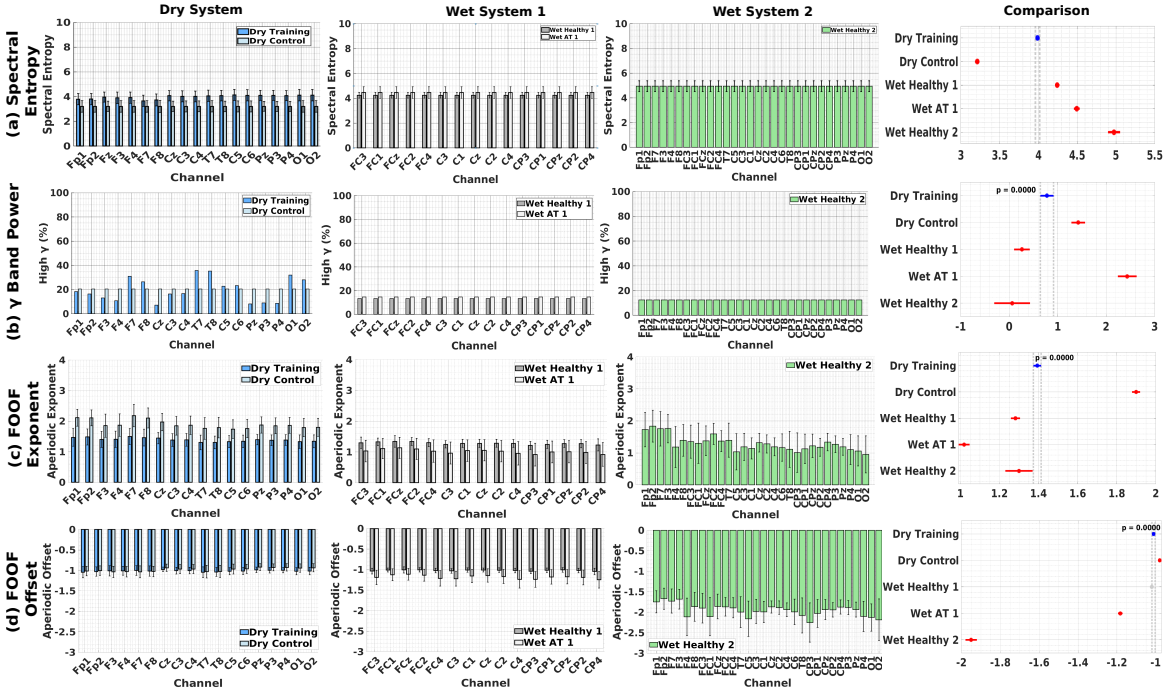

Figure 15. Different frequency-domain quality matrices after re-referencing to Fz for all five datasets. Panels show (a) spectral entropy, (b) γ-band power, (c) aperiodic exponent, and (d) Fitting Oscillations and One-Over F (FOOF). The near-perfect overlap of native-reference and Fz-referenced results confirms that frequency-based metrics are similarly robust to the choice of reference electrode.

observed in γ-band power also support the notion that context-specific factors—such as participant movement, muscular tension, and environmental noise–critically downgrade the signal's characteristics, independent of the electrode type.

The central outcome of this analysis is again the confirmation that state-of-the-art dry-electrode EEG systems can indeed yield data quality comparable to conventional wet-electrode setups. Specifically, our results reveal that dry-electrode EEG recordings obtained in intensely noisy and dynamic environments (Dry Control) do experience heightened transient artifacts such as increased spike counts, elevated peak amplitudes, higher γ-band activity, and prominent broadband power as reflected in FOOF metrics. However, dry-electrode recordings captured under moderately noisy conditions (Dry Training) exhibit artifact profiles approaching those observed in wet electrode recordings conducted under controlled laboratory or clinical conditions. This suggests that, in EEG electrode manufacturing technology, the electrode type alone does not crucially limit EEG quality; rather, the environmental conditions emerge as the most prominent factor impacting the reliability of acquired EEG.

Along these lines, the investigation on the wet-electrode datasets (Wet Healthy 1, Wet AT 1, and Wet Healthy 2) demonstrates that wet electrodes and, even, controlled laboratory environments do not guarantee artifact-free recordings, with consistent evidence of frontal-central contamination presumably arising from ocular and facial muscle artifacts. This frontal (and, more generally, peripheral) channel vulnerability of wet systems contrasts with the dry-electrode recording susceptibility to widespread temporal and central noise profiles likely arising due to external interference and participant movements. Hence, there currently seems to be a less clear overall dominance of wet sensors over dry ones, where both categories of sensors seem to have their particular vulnerabilities, pros and cons.

ICLabel analysis further confirmed that the composition of artifact types and their severity differed substantially across EEG systems and recording environments. Although the proportion of brain-related components remained relatively stable between the dry-electrode datasets (Dry Training and Dry Control), the noisier public exhibition environment substantially amplified movement-related and mixed-source artifacts. Conversely, laboratory-based wet EEG recordings (Wet Healthy 1 and 2) demonstrated the cleanest IC profiles, emphasizing the advantage of controlled conditions in reducing artifact contamination. Semi-controlled conditions (Wet AT 1) also presented moderate artifact contamination, highlighting that environmental factors and participant activities substantially influence artifact distributions, regardless of electrode technology. Collectively, these findings reinforce

the importance of context-specific artifact mitigation strategies when employing EEG systems in varied operational settings. A like-for-like literature benchmark for ICLabel proportions in large, in-the-wild SMR BCI datasets is currently lacking. Our observed 31–49.5% Brain range is, however, in line with literature reporting higher Brain shares (around 50%) in clean, resting-state data and approximately 30% Brain components in other task and environment contexts, reinforcing the opinion that the recording context drives the observed class mix of ICLabel; critically, increases in γ-power and peak amplitude corroborate heavier contamination in Dry Control. [44, 46, 47].

It must be further noted that, with the exception of the ICLabel analysis, where specific artifact sources can be explicitly identified, we have deliberately (mostly) avoided associating the examined quality metrics (and the corresponding EEG system performance) with specific artifacts. The reason is that, despite some relevant literature [31, 48] and existing sound hypotheses (e.g., spikes in frontal channels that can be relatively safely attributed more often than not to eye blinks and movements), the exact EEG signatures of various artifact sources are not very well established yet. However, as this line of research continuously develops, we hope that the detailed metric benchmarking offered in this manuscript will serve in the future as a go-to resource to predict the suitability or contraindication of a given EEG sensor type for specific recording environments and mental tasks.

4.3. Limitations and Confounds

While this study offers a systematic and multifaceted comparison of dry and wet EEG systems across diverse recording environments, it is important to acknowledge certain important limitations and confounds. The datasets analyzed were not fully matched, most critically, in terms of experimental conditions, but also with regard to the number of subjects and the participants' characteristics (especially, the distinction able-bodied vs. with motor disabilities).

Dataset Heterogeneity: Starting with what we view as the most influential confounding factor, namely, the ensemble of environmental conditions (likelihood and intensity of noise in the surroundings) and task demands (open- vs. closed-loop MI, control of visual feedback or an actual device, performing MI with a surrounding crowd or not, etc.), it is clear that the respective conditions differed substantially across datasets, ranging from dry EEG-based "machine control" sessions to wet open-loop calibration runs. It is forthrightly acknowledged throughout our investigation that the five analyzed dataset groups (Dry Control, Dry Training, Wet Healthy 1, Wet AT 1, Wet Healthy 2) cannot make a clear distinction between dry and wet sensors per se, but rather represent combinations of sensor technology type (dry vs. wet) and environmental conditions with increasing (albeit, not strictly measurable) amounts of noise and artifacts, in order: Wet Healthy 2, Wet Healthy 1, Wet AT 1, Dry Training, Dry Control. We have assumed that Wet Healthy 1 recordings had somewhat higher chances of being affected by noise compared to Wet Healthy 2, as, despite being recorded in a lab environment, the majority of these served to prep users for BCI application control [25] and often took place in open university spaces, under tight and hectic schedules, and/or with several operators present; conversely, Wet Healthy 2 data have been recorded with stricter experimental protocols in the lab [26]. Furthermore, artifact contamination in the Wet AT 1 dataset (obtained with the gUSBamp wet system like Wet Healthy 1) is believed to be higher than in Wet Healthy 1 as the data were recorded in much less controlled and hectic environments in the presence of other electronic devices (clinic, end-user homes), while also several participants of this group suffered from artifact-inducing conditions such as spasms. Dry Training and Control, as already explained, were recorded in the comparatively harshest conditions in the Mental Work exhibition premises. However, Dry Training took place in a loosely isolated booth with a higher degree of privacy that must have considerably mitigated the exhibition's ambient noise and the vulnerability to self-induced physiological artifacts. Consequently, we have considered that Wet AT 1 and Dry Training correspond to similar levels of artifact and noise contamination, while Dry Control undoubtedly constitutes, by far, the toughest condition in that regard. These assumptions comply very well with our findings, taking into account all aspects and metrics of our analysis, notwithstanding that the order of signal quality is, in the case of a few metrics, slightly altered. Overall, in spite of the confounding of sensor type and environmental noise in our data, we strongly argue that the profound similarities between Dry Training and Wet AT 1 with respect to noise vulnerability, as well as the availability of data across the aforementioned large spectrum of potential noise and artifact sources that allows meaningful interpolations, give credibility to our main claim that state-of-the-art dry sensors are nowadays competitive to wet electrodes in real-world scenarios. We pinpoint the fact that such confounding was inevitable, since we did not have the chance to design a dedicated experiment but, rather, resorted to leveraging historical data and combine them in order to collect evidence towards answering an important research question. We postulate that what this study suffers from due to the surrounding noise confound, it makes up thanks to the

sheer amount of data we were able to accumulate, going well beyond the state-of-the-art in that respect.

Sample Size Difference: It must be additionally highlighted that the total sample sizes dry vs. wet differ substantially (530 dry vs. 71 + 80 = 151 wet). Nevertheless, this imbalance is largely countered by the fact that wet recordings averaged more and longer runs/sessions per participant, so that the total available time of EEG data is approximately balanced. Importantly, the sample size of the minority group (wet, 151 participants) is anyway big enough to yield reliable statistics. The recording hardware and sampling rate also varied: dry data were acquired at 300 Hz across 19 channels; Wet Healthy 1 and Wet AT at 512 Hz over 16 channels; and Wet Healthy 2 at 1000 Hz with 119 channels (only the subset common with the other systems was analyzed, though). However, by holding channel overlap constant, offering direct comparisons on the four channels that were common in all three layouts and by applying identical quality-metric definitions, we posit that these differences had no impact whatsoever on the conclusions reached here.

Reference Montage Effects: Another potential confounding factor could be the use of different reference electrode location across datasets. In the dry EEG system, channels were referenced to the average of the two earlobe potentials. In Wet Healthy 1 and Wet AT, the right earlobe was used as reference, while in Wet Healthy 2, the reference electrode was positioned at the nasion. We denote that all three reference choices (which are all standard choices in the field), despite not coinciding with one another, are on neural-activity-neutral sites and, thus, interchangeable. Additionally, they are all far enough from the channels of interest so that their effect on the recorded activity must be negligible, if at all present. Most crucially, the vast majority of the quality metrics used in this study—kurtosis, spiking activity, spectral entropy, AMI, and FOOF parameters—are shape- or distribution-based and, therefore, amplitude-invariant and robust to referencing differences. We only included a few amplitude-sensitive metrics (such as standard deviation and peak amplitude), which, in theory, may have very slightly attenuated amplitudes on the nearby peripheral electrode sites. In practice, dedicated inspection confirmed that peripheral channels exhibit similar EEG amplitude distributions in all datasets. Last but not least, when considering re-referencing to a commonly available channel (to avoid any debate over the influence of the reference), one must take into account that a non-optimal site could introduce additional artifacts or distortions, and/or lead to losing significant information. Specifically, among the available options (C3, C4, Cz, Fz) we exclude the lateral electrodes C3 and C4, so that the only remaining options are Cz and Fz. Cz is an active site for motor-related neural activity and may not offer a neutral baseline. The only other potential alternative, Fz, is also problematic due to its susceptibility to ocular and facial muscle artifacts (Fig. 8 supports that Fz is susceptible to noise); re-referencing to Fz could inadvertently spread noise across all electrodes, thereby compromising the integrity of spatial signal features. For all these reasons, we opted to retain each dataset's native reference configuration when presenting our main results. However, for the sake of completeness, we verified that re-referencing to Fz would not significantly alter our findings and conclusions. We re-computed all metrics after applying a common Fz re-referencing across every dataset. As shown in Fig. 14 and Fig. 15, the resulting time- and frequency-domain measures remained virtually unchanged. Only standard deviation plots showed limited, localized differences, confirming that our quality metrics are robust to the choice of reference, even when re-referencing to an electrode prone to ocular and facial artifacts (Fz).

4.4. Practical Implications and Future Directions

We wish to underline that this study does not aim and cannot be used to assess specific products and manufacturers. Given the discussion on confounds above, it is clearly acknowledged that a strict statistical comparison of systems is scientifically unattainable with the data considered here. Instead, we use data from these three commercial and popular systems (one dry, two wet) in the EEG community opportunistically, solely because the authors happened to have access to a big amount of data from these particular products. Considering that all these systems are competitive, well-known and widely-used products in the BCI industry, we take the assumption that they can represent the respective state-of-the-art, and hence this study assesses the current competencies of dry and wet EEG technologies. However, by no means do we imply that these are the sole products/manufacturers that could potentially be employed for benchmarking, or attempt direct system-to-system comparisons, and we encourage other researchers to perform similar analyses using other EEG recording devices.

Concluding, the comparative analysis presented here confirms the practicality and effectiveness of dry-electrode EEG systems, even when deployed in highly challenging settings, such as public exhibitions or when involving dynamic, real-world interaction. Although specific channel vulnerabilities and artifact patterns differ between electrode technologies, the overall

data quality attainable by contemporary dry-electrode systems proves sufficiently robust to match wet-electrode benchmarks, especially provided appropriate signal processing, artifact mitigation and online adaptation [49, 50, 51] strategies are implemented. Future developments could further enhance dry EEG reliability by refining electrode designs and developing sophisticated artifact mitigation algorithms tailored explicitly to dynamic environmental conditions. Such advances will ensure that dry EEG remains not only a viable alternative but a preferred solution for mobile, user-friendly, and scalable neuroimaging applications in naturalistic and real-world scenarios.

Acknowledgments

The authors would like to thank the Mental Work Team, including Laurent Bolli, Michael Mitchell, Jonathon Keats, and Arnaud Desvachez, for developing the Mental Work Research Platform, which enabled users to interact with neural interfaces through the existing Mental Work system and provided a large volume of valuable brain signal data that was instrumental in the preparation of this manuscript. The Mental Work Exhibition was supported by the Media Engineering Institute (HEIG-VD), École Polytechnique Fédérale de Lausanne (EPFL), Fondation Campus Biotech Genève, the Human Neuroscience Platform, the Swiss National Science Foundation, the Hasler Foundation, the Gebert Rüf Foundation, and the Ernst Göhner Foundation.

References

[1] Hinrichs H, Scholz M, Baum A K, Kam J W Y, Knight R T and Heinze H J 2020 *Scientific Reports* **10** 62154
[2] Liao L D, Chen C Y, Wang I J, Chen S F, Li S Y, Chen B W, Chang J Y and Lin C T 2012 *Journal of NeuroEngineering and Rehabilitation* **9** 1–12
[3] Gargiulo G, Calvo R A, Bifulco P, Cesarelli M, Jin C, Mohamed A and van Schaik A 2010 *Clinical Neurophysiology* **121** 686–693
[4] Guger C, Krausz G, Allison B Z and Edlinger G 2012 *Frontiers in Neuroscience* **6**
[5] De Vos M, Kroesen M, Emkes R and Debener S 2014 *Journal of Neural Engineering* **11** 036008
[6] Yeung A, Garudadri H, Van Toen C, Mercier P, Balkan O, Makeig S and Virji-Babul N 2015 Comparison of foam-based and spring-loaded dry eeg electrodes with wet electrodes in resting and moving conditions *Proceedings of the 2015 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp 7131–7134
[7] Saab J, Battes B and Grosse-Wentrup M 2011 Simultaneous eeg recordings with dry and wet electrodes in motor-imagery *Proceedings of the 5th International Brain–Computer Interface Conference (BCI 2011)* (Verlag der Technischen Universität Graz) pp 312–315
[8] Fiedler P, Haueisen J, Jannek D, Griebel S, Zentner L, Vaz F and Fonseca C 2014 *Acta IMEKO* **3** 33–37
[9] Mathewson K E, Harrison T J L and Kizuk S A D 2017 *Psychophysiology* **54** 74–82
[10] Kamousi B, Grant A M, Bachelder B, Yi J, Hajinoroozi M and Woo R 2019 *Clinical Neurophysiology Practice* **4** 69–75
[11] Wang C H, Moreau D and Kao S C 2019 *Frontiers in Neuroscience* **13** 893
[12] Pinho F, Cerqueira J, Correia J, Sousa N and Dias N 2017 *Journal of Medical Engineering & Technology* **41** 564–585
[13] Doerrfuss J I, Kilic T, Ahmadi M, Weber J E and Holtkamp M 2020 *Epilepsy & Behavior* **104** 106486
[14] Lim C G, Lee T S, Guan C, Fung D S S, Zhao Y, Teng S S W, Zhang H and Krishnan K R R 2012 *PLOS ONE* **7** e46692
[15] Wang H, Dragomir A, Abbasi N I, Li J, Thakor N V and Bezerianos A 2018 *Cognitive Neurodynamics* **12** 365–376
[16] So W K Y, Wong S W H, Mak J N and Chan R H M 2017 *PLOS ONE* **12** e0174949
[17] Croft R J and Barry R J 2000 *Neurophysiologie Clinique / Clinical Neurophysiology* **30** 5–19
[18] Schlögl A, Keinrath C, Zimmermann D, Scherer R, Leeb R and Pfurtscheller G 2007 *Clinical Neurophysiology* **118** 98–104
[19] Islam M K, Rastegarnia A and Yang Z 2016 *Neurophysiologie Clinique / Clinical Neurophysiology* **46** 287–305
[20] Jin Z, Bourban F, Leeb R and Perdikis S 2022 Quantifying the impact and profiling functional eeg artifacts *Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* pp 629–634
[21] Matran-Fernandez A, Valeriani D and Poli R 2017 Toward bcis out of the lab: Impact of motion artifacts on brain–computer interface performance *Wireless Medical Systems and Algorithms* (CRC Press) pp 219–239
[22] Jiang X, Bian G B and Tian Z 2019 *Sensors* **19** 987
[23] Mental work: The cognitive revolution starts here `https://mentalwork.net/project/` accessed: 16 December 2025
[24] Sultana M and Perdikis S 2024 *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **32** 3422–3431
[25] Leeb R, Perdikis S, Tonin L, Biasiucci A, Tavella M, Creatura M, Molina A, Al-Khodairy A, Carlson T and Millán J d R 2013 *Artificial Intelligence in Medicine* **59** 121–132
[26] Blankertz B, Sannelli C, Halder S, Hammer E M, Kübler A, Müller K R, Curio G and Dickhaus T 2010 *NeuroImage* **51** 1303–1309
[27] Sannelli C, Vidaurre C, Müller K R and Blankertz B 2019 *PloS one* **14** e0207351
[28] Welch P 1967 *IEEE Transactions on Audio and Electroacoustics* **15** 70–73
[29] Perdikis S, Tonin L, Saeedi S, Schneider C and Millán J d R 2018 *PLOS Biology* **16** e2003787
[30] Müller-Putz G, Scherer R, Brunner C, Leeb R and Pfurtscheller G 2008 *International Journal of Bioelectromagnetism* **10** 52–55
[31] Daly I, Scherer R, Billinger M and Müller-Putz G 2014 *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **23** 725–736
[32] Nicolaou N and Nasuto S J 2007 *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* **48** 173–183
[33] Talsma D 2008 *Psychophysiology* **45** 216–228
[34] Mahajan R and Morshed B I 2014 *IEEE Journal of Biomedical and Health Informatics* **19** 158–165
[35] Delorme A, Sejnowski T and Makeig S 2007 *NeuroImage* **34** 1443–1449
[36] Nolan H, Whelan R and Reilly R B 2010 *Journal of Neuroscience Methods* **192** 152–162

[37] Sagha H, Perdikis S, Millán J d R and Chavarriaga R 2015 *IEEE Transactions on Biomedical Engineering* **62** 858–864

[38] Mantini D, Perrucci M G, Del Gratta C, Romani G L and Corbetta M 2007 *Proceedings of the National Academy of Sciences* **104** 13170–13175

[39] Yong X, Ward R K and Birch G E 2008 Robust common spatial patterns for eeg signal preprocessing *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp 2087–2090

[40] Niedermeyer E *et al.* 2005 The normal eeg of the waking adult *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (Lippincott Williams & Wilkins) pp 155–164

[41] Goncharova I I, McFarland D J, Vaughan T M and Wolpaw J R 2003 *Clinical Neurophysiology* **114** 1580–1593

[42] Donoghue T, Haller M, Peterson E J, Varma P, Sebastian P, Gao R, Noto T, Lara A H, Wallis J D, Knight R T *et al.* 2020 *Nature Neuroscience* **23** 1655–1665

[43] Lee T W, Girolami M and Sejnowski T J 1999 *Neural Computation* **11** 417–441

[44] Pion-Tonachini L, Kreutz-Delgado K and Makeig S 2019 *NeuroImage* **198** 181–197

[45] Guger C, Edlinger G, Harkam W, Niedermayer I and Pfurtscheller G 2003 *IEEE transactions on neural systems and rehabilitation engineering* **11** 145–147

[46] SCCN/EEGLAB 2020 The eeglab news #3: Iclabel q&a accessed 15 Oct 2025 URL `https://sccn.ucsd.edu/eeglab/eeglab_news/3/ICLabel_Q%26A.php`

[47] ICLabel Team, SCCN 2025 Iclabel tutorial: Eeg independent component labeling accessed 15 Oct 2025 URL `https://labeling.ucsd.edu/tutorial/labels`

[48] Jin Z, Bourban F, Leeb R and Perdikis S 2022 Quantifying the impact and profiling functional eeg artifacts *Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* pp 629–634

[49] Perdikis S, Leeb R and Millán J d R 2016 *Journal of Neural Engineering* **13** 036018

[50] Perdikis S, Leeb R, Chavarriaga R and Millán J d R 2021 *IEEE Transactions on Neural Networks and Learning Systems* **32** 3471–3483

[51] Cunha J D, Perdikis S, Halder S and Scherer R 2021 *IEEE Access* **9** 41688–41703